SIAM J. MATH. ANAL.
Vol. 32, No. 1, pp. 1–29

SCALING VARIABLES AND STABILITY OF HYPERBOLIC FRONTS*

THIERRY GALLAY† AND GENEVIÈVE RAUGEL†

Abstract. We consider the damped hyperbolic equation

$$\varepsilon u_{tt} + u_t = u_{xx} + \mathcal{F}(u) , \quad x \in \mathbf{R} , \quad t \geq 0 , \tag{1}$$

where ε is a positive, not necessarily small parameter. We assume that $\mathcal{F}(0) = \mathcal{F}(1) = 0$ and that $\mathcal{F}$ is concave on the interval $[0, 1]$. Under these hypotheses, (1) has a family of monotone traveling wave solutions (or propagating fronts) connecting the equilibria $u = 0$ and $u = 1$. This family is indexed by a parameter $c \geq c_*$ related to the speed of the front. In the critical case $c = c_*$, we prove that the traveling wave is asymptotically stable with respect to perturbations in a weighted Sobolev space. In addition, we show that the perturbations decay to zero like $t^{-3/2}$ as $t \to +\infty$ and approach a universal self-similar profile, which is independent of ε, F, and the initial data. In particular, our solutions behave for large times like those of the parabolic equation obtained by setting $\varepsilon = 0$ in (1). The proof of our results relies on various energy estimates for (1) rewritten in self-similar variables $x/\sqrt{t}$, $\log t$.

Key words. damped wave equation, traveling wave, stability, asymptotic behavior, self-similar variables

AMS subject classifications. 35B40, 35B35, 35B30, 35L05, 35C20

PII. S0036141099351334

1. Introduction.

In this paper, we study the asymptotic stability of traveling wave solutions to nonlinear damped hyperbolic equations on the real line. Besides describing the propagation of voltage along nonlinear transmission lines, these equations have been proposed as mathematical models for spreading and interacting particles [DO, Ha2, Ha3]. In the latter context, they provide an alternative to the reaction-diffusion systems which are very common in chemistry and biology, especially in genetics and population dynamics [Mu]. The two classes of models differ by the choice of the stochastic process describing the spatial spread of the individuals: instead of Brownian motion, the damped hyperbolic equations are based on a more realistic velocity jump process which takes into account the inertia of the particles [Go, Kac, Za]. Since this process is asymptotically diffusive, the long-time behavior of the solutions is expected to be essentially parabolic [GR2].

We study here the simple case of a scalar equation with a "monostable" nonlinearity. To be specific, we consider the equation

$$\varepsilon U_{TT} + U_T = U_{XX} + \mathcal{F}(U) , \tag{1.1}$$

where $X \in \mathbf{R}$, $T \geq 0$, and ε is a positive, *not necessarily small* parameter. We assume that the nonlinearity $\mathcal{F} \in \mathcal{C}^2(\mathbf{R})$ satisfies

$$\mathcal{F}(0) = \mathcal{F}(1) = 0 , \quad \mathcal{F}'(0) > 0 , \quad \mathcal{F}'(1) < 0 , \quad \mathcal{F}''(U) \leq 0 \ \text{ for } \ U \in [0, 1] . \tag{1.2}$$

In particular, $U = 1$ is a stable equilibrium of (1.1), and $U = 0$ is unstable. A typical nonlinearity satisfying (1.2) is $\mathcal{F}(U) = U - U^m$, with $m \in \mathbf{N}$, $m \geq 2$.

*Received by the editors February 5, 1999; accepted for publication (in revised form) September 22, 1999; published electronically June 22, 2000.
http://www.siam.org/journals/sima/32-1/35133.html

†Laboratoire de Mathématique, Université de Paris-Sud, UMR 8628 du CNRS, Bâtiment 425, F-91405 Orsay, France (Thierry.Gallay@math.u-psud.fr, Genevieve.Raugel@math.u-psud.fr).

Under the assumptions in (1.2), equation (1.1) has monotone traveling wave solutions (or propagating fronts) connecting the equilibrium states $U = 1$ and $U = 0$ [Ha1, GR1]. Indeed, choosing $c > 0$ and setting $U(X,T) = h(\sqrt{1+\varepsilon c^2}X - cT)$, we obtain for h the ordinary differential equation

$$h''(\xi) + ch'(\xi) + \mathcal{F}(h(\xi)) \,=\, 0 \;, \quad \xi \in \mathbf{R} \;. \tag{1.3}$$

Equation (1.3) is known to have a strictly decreasing solution satisfying $h(-\infty) = 1$ and $h(+\infty) = 0$ if and only if $c \geq c_* = 2\sqrt{\mathcal{F}'(0)}$ [KPP, AW]. This solution is unique up to translations in the variable ξ. Thus, (1.1) has a family of monotone traveling waves indexed by the speed parameter $c \geq c_*$. Note that the actual speed of the wave is not c but $c/\sqrt{1+\varepsilon c^2}$, a quantity which is bounded by $1/\sqrt{\varepsilon}$ for all $c \geq c_*$.

In an earlier paper [GR1], we investigated the stability of the traveling waves of (1.1) in the case where $\mathcal{F}(U) = U - U^2$. In particular, we showed that for all $\varepsilon > 0$ and all $c \geq c_*$, the front h is asymptotically stable with respect to small perturbations in a weighted Sobolev space (with exponential weight). This local stability result holds, in fact, for all nonlinearities satisfying (1.2); see [GR3]. In addition, if $\varepsilon > 0$ is sufficiently small, we proved in [GR1] that the front h is stable with respect to large perturbations, provided some positivity conditions are fulfilled. This global stability property relies on the hyperbolic maximum principle and can also be extended to more general nonlinearities [GR3]. Finally, we showed in all cases that the perturbations converge uniformly to zero faster than $T^{-1/4}$ as $T \to +\infty$.

When $\varepsilon \to 0$, (1.1) reduces to the semilinear parabolic equation $U_T = U_{XX} + \mathcal{F}(U)$, which has been intensively studied since the pioneering works of Fisher [Fi] and Kolmogorov, Petrovskii, and Piskunov [KPP]. In particular, the parabolic maximum principle and probabilistic techniques have been used to show the convergence of a large class of solutions to traveling waves [AW, Br]. In the more general context of parabolic systems, a local stability analysis of the waves has been initiated by Sattinger [Sa] and extended by many authors [Ki, EW, Kap, BK1, RK], using resolvent estimates, energy functionals, and renormalization techniques. Finally, in the critical case $c = c_*$, it has been proved by one of us [Ga] that the perturbations of the front decay to zero like $T^{-3/2}$ as $T \to +\infty$ and approach a universal self-similar profile. The aim of the present paper is precisely to extend this detailed convergence result to the hyperbolic case $\varepsilon > 0$. Together with earlier results from [GR1, GR3], this will provide a fairly complete picture of the stability properties of the traveling waves of (1.1).

To study the stability of the critical front h with $c = c_*$, it is convenient to go to a moving frame using the change of variables $U(X,T) = V(\sqrt{1+\varepsilon c_*^2}X - c_*T, T)$. The equation for V is

$$\varepsilon V_{TT} + V_T - 2\varepsilon c_* V_{\xi T} \,=\, V_{\xi\xi} + c_* V_\xi + \mathcal{F}(V) \;, \tag{1.4}$$

where $\xi = \sqrt{1+\varepsilon c_*^2}X - c_*T$. By construction, h is a stationary solution of (1.4). Following [Ki, Ga], we consider perturbed solutions of the form

$$V(\xi,T) \,=\, h(\xi) + w(\xi,T) \,\equiv\, h(\xi) + h'(\xi) W\left(\xi, \frac{T}{1+\varepsilon c_*^2}\right) \;. \tag{1.5}$$

The equation satisfied by the perturbation w is

$$\varepsilon w_{TT} + w_T - 2\varepsilon c_* w_{\xi T} \,=\, w_{\xi\xi} + c_* w_\xi + \mathcal{F}'(h) w + \mathcal{N}(h,w) w^2 \;, \tag{1.6}$$

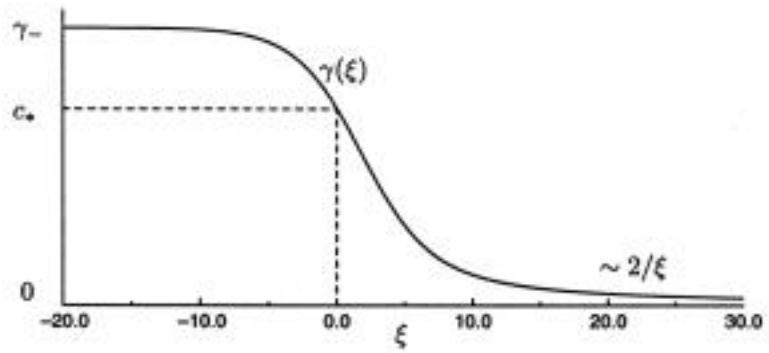


FIG. 1. *The function $\gamma(\xi)$ in the case where $\mathcal{F}(U) = U - U^2$ (hence $c_* = 2$, $\gamma_- = 2\sqrt{2}$).*

where

$$\mathcal{N}(a,b) \,=\, \int_0^1 (1-s)\mathcal{F}''(a+sb)\,\mathrm{d}s \,=\, \frac{1}{b^2}\big(\mathcal{F}(a+b) - \mathcal{F}(a) - b\mathcal{F}'(a)\big)\ . \tag{1.7}$$

The Ansatz $w(\xi,T) = h'(\xi)W(\xi,\tau)$, where $\tau = T/(1+\varepsilon c_*^2)$, is motivated by the fact that $W(\xi,\tau)$ becomes asymptotically self-similar as $T \to +\infty$, while the actual perturbation $w(\xi,T)$ does not; see Corollary 1.3 below. We remark that this definition makes sense, since $h'(\xi) < 0$ for all $\xi \in \mathbf{R}$. The equation for W reads

$$\eta W_{\tau\tau} + (1-\nu\gamma(\xi))W_\tau - 2\nu W_{\xi\tau} \,=\, W_{\xi\xi} + \gamma(\xi)W_\xi + h'(\xi)W^2\mathcal{N}(h(\xi),h'(\xi)W)\ , \tag{1.8}$$

where

$$\eta \,=\, \frac{\varepsilon}{(1+\varepsilon c_*^2)^2}\ , \quad \nu \,=\, \frac{\varepsilon c_*}{1+\varepsilon c_*^2}\ , \quad \gamma(\xi) \,=\, c_* + 2\frac{h''(\xi)}{h'(\xi)}\ . \tag{1.9}$$

Before analyzing the solutions of (1.8), we briefly comment on the definitions (1.9). We first remark that there is no loss of generality in assuming $\varepsilon = 1$ in (1.1), since $(\varepsilon,\mathcal{F})$ can be transformed into $(1,\varepsilon\mathcal{F})$ by rescaling X and T. However, we find it more convenient to fix the nonlinearity $\mathcal{F}$ and to consider ε as a free parameter. Then $c_* > 0$ is fixed, and η,ν are functions of ε only. These expressions are not independent, since $\nu^2 + \eta = \nu/c_*$. Observe also that η,ν are uniformly bounded for all $\varepsilon > 0$, and converge to zero as $\varepsilon \to 0$. We now list the properties of the "drift" $\gamma(\xi)$ which will be crucial for our analysis. From [Sa, AW], we know that the front h (with $c = c_*$) satisfies

$$h(\xi) \,=\, \begin{cases} 1 - a_3\mathrm{e}^{\kappa\xi} + \mathcal{O}(\mathrm{e}^{2\kappa\xi}) & \text{as } \xi \to -\infty\ , \\ (a_1\xi + a_2)\mathrm{e}^{-c_*\xi/2} + \mathcal{O}(\xi^2\mathrm{e}^{-c_*\xi}) & \text{as } \xi \to +\infty\ , \end{cases} \tag{1.10}$$

where $a_1, a_3 > 0$, $a_2 \in \mathbf{R}$, and $\kappa = \frac{1}{2}(-c_* + \sqrt{c_*^2 - 4\mathcal{F}'(1)}) > 0$. Using (1.10) and similar asymptotic expansions for the derivatives h', h'', we obtain

$$\gamma(\xi) \,=\, \begin{cases} \gamma_- + \mathcal{O}(\mathrm{e}^{\kappa\xi}) & \text{as } \xi \to -\infty\ , \\ 2/(\xi+\xi_0) + \mathcal{O}(\xi\mathrm{e}^{-c_*\xi/2}) & \text{as } \xi \to +\infty\ , \end{cases} \tag{1.11}$$

where $\gamma_- = c_* + 2\kappa = 2\sqrt{\mathcal{F}'(0) - \mathcal{F}'(1)}$ and $\xi_0 = (a_2/a_1 - 2/c_*)$. It also follows from (1.3), (1.9) that

$$\gamma'(\xi) \,=\, -\frac{1}{2}\gamma(\xi)^2 + 2\big(\mathcal{F}'(0) - \mathcal{F}'(h(\xi))\big)\ , \quad \xi \in \mathbf{R}\ . \tag{1.12}$$

Together with (1.2), this equation implies that $-\frac{1}{2}\gamma(\xi)^2 \le \gamma'(\xi) \le 0$ for all $\xi \in \mathbf{R}$. Indeed, the lower bound on $\gamma'(\xi)$ is obvious, and the upper bound follows from the inequality $\gamma''(\xi) + \gamma(\xi)\gamma'(\xi) \le 0$ obtained by differentiating (1.12). In fact, we even have $\gamma'(\xi) < 0$ whenever $\gamma(\xi) < \gamma_-$. Replacing $h(\xi)$ by a translate, we may (and will always) assume that $\gamma(0) = c_*$, i.e., $h''(0) = 0$; see Figure 1. This amounts to fixing the origin in the moving frame.

To study the behavior of the solutions W of (1.8), we use the *scaling variables* or *self-similar variables* defined by

$$x = \frac{\xi}{\sqrt{\tau+\tau_0}}\ , \quad t = \log(\tau+\tau_0)\ , \tag{1.13}$$

where $\tau_0 \ge 1$ will be fixed later. These variables have been widely used to investigate the long-time behavior of solutions to parabolic equations, in particular, to prove convergence to self-similar solutions [Kav, EZ, GV, EKM, BK2, Wa, GM]. Although the scaling (1.13) is parabolic in essence, we have shown in [GR2] that self-similar variables are also a powerful tool in the realm of damped hyperbolic equations. The reason is that the long-time behavior of such systems is often determined by simpler parabolic equations; see [HL, Ni, GR2] for specific examples of this phenomenon. In our case, the result of [Ga] in the parabolic limit $\varepsilon = 0$ suggests that $W(\xi,\tau)$ should behave like $\tau^{-3/2}\varphi^*(\xi/\sqrt{\tau})$ as $\tau \to +\infty$, where φ^* is given by (1.21) below. Thus, following the method developed in [GR2], we define rescaled functions u and v by

$$u(x,t) = \mathrm{e}^{3t/2}W(x\mathrm{e}^{t/2}, \mathrm{e}^t - \tau_0)\ , \quad v(x,t) = \mathrm{e}^{5t/2}W_\tau(x\mathrm{e}^{t/2}, \mathrm{e}^t - \tau_0)\ , \tag{1.14}$$

or equivalently

$$\begin{aligned} W(\xi,\tau) &= \frac{1}{(\tau+\tau_0)^{3/2}}\, u\left(\frac{\xi}{\sqrt{\tau+\tau_0}}\,, \log(\tau+\tau_0)\right)\ , \\ W_\tau(\xi,\tau) &= \frac{1}{(\tau+\tau_0)^{5/2}}\, v\left(\frac{\xi}{\sqrt{\tau+\tau_0}}\,, \log(\tau+\tau_0)\right)\ . \end{aligned} \tag{1.15}$$

Then the functions $u(x,t), v(x,t)$ satisfy the system

$$\begin{aligned} &u_t - \frac{x}{2}u_x - \frac{3}{2}u = v\ , \\ &\eta\mathrm{e}^{-t}\left(v_t - \frac{x}{2}v_x - \frac{5}{2}v\right) + (1 - \nu\gamma(x\mathrm{e}^{t/2}))v - 2\nu\mathrm{e}^{-t/2}v_x \\ &= u_{xx} + \mathrm{e}^{t/2}\gamma(x\mathrm{e}^{t/2})u_x + \mathrm{e}^{-t/2}h'(x\mathrm{e}^{t/2})u(x,t)^2N(x,t)\ , \end{aligned} \tag{1.16}$$

where $x \in \mathbf{R}$, $t \ge t_0 = \log\tau_0$, and $N(x,t) = \mathcal{N}(h(x\mathrm{e}^{t/2}), \mathrm{e}^{-3t/2}h'(x\mathrm{e}^{t/2})u(x,t))$.

We next introduce the function spaces in which we shall study the solutions of (1.16). For $t \ge 0$, $k \in \mathbf{N}$, we denote by L^2_t, H^k_t the weighted Lebesgue and Sobolev spaces defined by the norms

$$\begin{aligned} \|u\|^2_{\mathrm{L}^2_t} &= \int_{-\infty}^0 \mathrm{e}^{2\kappa x\mathrm{e}^{t/2}}u(x)^2\,\mathrm{d}x + \int_0^\infty (1+x)^6u(x)^2\,\mathrm{d}x\ , \\ \|u\|^2_{\mathrm{H}^k_t} &= \sum_{i=0}^k \|\partial_x^i u\|^2_{\mathrm{L}^2_t}\ , \end{aligned} \tag{1.17}$$

where κ appears in (1.10). Our basic space will be the product $\mathrm{Z}_t = \mathrm{H}^1_t \times \mathrm{L}^2_t$ equipped with the standard norm $\|(u,v)\|_{\mathrm{Z}_t} = (\|u\|^2_{\mathrm{H}^1_t} + \|v\|^2_{\mathrm{L}^2_t})^{1/2}$. In order to state results which are uniform in ε as $\varepsilon \to 0$, it is convenient to introduce also the quadratic form

$$\Phi_\eta(t,u,v) \,=\, \|u\|^2_{\mathrm{H}^1_t} + \eta \mathrm{e}^{-t} \|v\|^2_{\mathrm{L}^2_t} \,. \tag{1.18}$$

From (1.14), (1.15), we see that $(u,v) \in \mathrm{Z}_t$ if and only if $(W, W_\tau) \in \mathrm{Z}_0 = \mathrm{H}^1_0 \times \mathrm{L}^2_0$. Moreover, since $h', h'' = \mathcal{O}(\mathrm{e}^{\kappa\xi})$ as $\xi \to -\infty$ and $h', h'' = \mathcal{O}(\xi \mathrm{e}^{-c_*\xi/2})$ as $\xi \to +\infty$, it is easy to verify that $(W, W_\tau) \in \mathrm{Z}_0$ if and only if the actual perturbation $w = h'W$ satisfies $(w, w_T) \in Y = Y^1 \times Y^0$, where the spaces Y^0, Y^1 are defined by the norms

$$\|w\|^2_{Y^0} \,=\, \int_{-\infty}^0 w^2 \,\mathrm{d}\xi + \int_0^\infty (1+\xi)^4 \mathrm{e}^{c_*\xi} w^2 \,\mathrm{d}\xi \,, \quad \|w\|^2_{Y^1} \,=\, \|w\|^2_{Y^0} + \|w_\xi\|^2_{Y^0} \,. \tag{1.19}$$

The comparison of (1.10), (1.19) reveals that the perturbations we consider decay to zero slightly faster than the front h itself as $\xi \to +\infty$. This is a necessary condition for stability, because the equilibrium state $U = 0$ of (1.1) is linearly unstable [Sa]. In particular, small translations of the front h are *not* allowed as perturbations.

Since our function space Z_t depends on time, we have to specify what we mean by a "solution of (1.16) in Z_t." As the system (1.16) has been obtained from the simpler equation (1.8) through the change of variables (1.14), the following definition is very natural.

Definition 1.1. *Let $t_2 > t_1 \geq t_0$, and let $\tau_i = \mathrm{e}^{t_i} - \tau_0$ for $i = 1, 2$. We say that "$(u,v) \in \mathcal{C}([t_1,t_2], \mathrm{Z}_t)$ is a solution of the system* (1.16)*" if there exists a (mild) solution $(W, W_\tau) \in \mathcal{C}([\tau_1,\tau_2], \mathrm{Z}_0)$ of* (1.8) *such that the relations* (1.14), (1.15) *hold.*

We recall that a *mild* solution of a partial differential equation is a continuous solution of the associated integral equation; see [Pa, section 4.2]. According to Definition 1.1, if $(u,v) \in \mathcal{C}([t_1,t_2], \mathrm{Z}_t)$ is a solution of (1.16), then $(u(t), v(t)) \in \mathrm{Z}_t$ for all $t \in [t_1, t_2]$. However, the continuity of (u,v) with respect to t has to be understood as the continuity in Z_0 of the functions (W, W_τ) defined by (1.15). In Proposition 2.2 below, we shall show that the Cauchy problem for (1.16) in Z_t is locally well posed.

Before stating our main result, we explain its content in a heuristic way. Taking formally the limit $t \to +\infty$ in (1.16) and using (1.11), we see that u satisfies the linear parabolic equation

$$u_t \,=\, \mathcal{L}_\infty u \,\stackrel{\mathrm{def}}{=}\, u_{xx} + \left(\frac{x}{2} + \frac{2}{x}\right) u_x + \frac{3}{2} u \quad \text{if } x > 0 \,, \quad u_x \,=\, 0 \quad \text{if } x \leq 0 \,. \tag{1.20}$$

Therefore, it is reasonable to expect that the long-time behavior of the solutions of (1.16) is determined by the spectral properties of the operator $\mathcal{L}_\infty$ on $\mathbf{R}_+$, with Neumann boundary condition at $x = 0$. Now, as is easily verified, this limiting operator is just the image under the scaling (1.15) of the radially symmetric Laplace operator in three dimensions. Indeed, if u and W are related through (1.15), the equation $u_t = \mathcal{L}_\infty u$ is equivalent to $W_\tau = W_{\xi\xi} + (2/\xi) W_\xi$, $\xi > 0$. This crucial observation explains the factor $(\tau+\tau_0)^{-3/2}$ in (1.15) and allows us to compute exactly the spectrum of $\mathcal{L}_\infty$ in various function spaces; see [GR2, Appendix A]. For instance, in the space $\mathrm{H}^1(\mathbf{R}_+, (1+x)^6 \mathrm{d}x)$, the spectrum of $\mathcal{L}_\infty$ consists of one simple, isolated eigenvalue at $\lambda = 0$, and of one "continuous" spectrum filling the half-plane $\{\lambda \in \mathbf{C} \,|\, \mathrm{Re}\,\lambda \leq -1/4\}$. The eigenfunction corresponding to $\lambda = 0$ is the Gaussian $\mathrm{e}^{-x^2/4}$. Therefore, we expect that the solution $u(x,t)$ of (1.16) converges as $t \to +\infty$ to

$\alpha\varphi^*(x)$ for some $\alpha \in \mathbf{R}$, where

$$\varphi^*(x) \,=\, \frac{1}{\sqrt{4\pi}} \begin{cases} 1 & \text{if } x < 0 \ , \\ \mathrm{e}^{-x^2/4} & \text{if } x \geq 0 \ . \end{cases} \tag{1.21}$$

This function is normalized so that $\int_0^\infty x^2 \varphi^*(x)\,\mathrm{d}x = 1$. Since $v = u_t - \frac{x}{2}u_x - \frac{3}{2}u$, we also expect that $v(x,t)$ converges to $\alpha\psi^*(x)$, where

$$\psi^* \,=\, -\frac{x}{2}\varphi^*_x - \frac{3}{2}\varphi^* \ . \tag{1.22}$$

It is crucial to note that (1.20) is independent of ε: this explains why the solutions of (1.8), and hence of (1.1), behave for large times like those of the corresponding parabolic equations.

Our main result, below, shows that these heuristic considerations are indeed correct.

THEOREM 1.2. *Assume that the nonlinearity $\mathcal{F}$ satisfies (1.2), and let $\varepsilon > 0$. There exist $t_0 > 0$, $\delta_0 > 0$, and $C > 0$ such that, for all initial data $(u_0, v_0) \in \mathrm{Z}_{t_0}$ with $\Phi_\eta(t_0, u_0, v_0) \leq \delta_0^2$, the system (1.16) has a unique solution $(u, v) \in \mathcal{C}([t_0, +\infty), \mathrm{Z}_t)$ satisfying $(u(t_0), v(t_0)) = (u_0, v_0)$. In addition, there exists $\alpha^* \in \mathbf{R}$ such that, for all $t \geq t_0$,*

$$\begin{aligned} \|u(t) - \alpha^*\varphi^*\|^2_{\mathrm{H}^1_t} + \eta\mathrm{e}^{-t}\|v(t) - \alpha^*\psi^*\|^2_{\mathrm{L}^2_t} + \int_{t_0}^t \mathrm{e}^{-(t-s)/2}\|v(s) - \alpha^*\psi^*\|^2_{\mathrm{L}^2_s}\,\mathrm{d}s \\ \leq\, C(1+t)^2\mathrm{e}^{-t/2}\Phi_\eta(t_0, u_0, v_0) \ . \end{aligned} \tag{1.23}$$

Remarks.

(1) In the proof of Theorem 1.2, we shall, for the sake of convenience, take the parameter $t_0 = \log(\tau_0)$ large enough, but this choice is irrelevant since, as reflected in Corollary 1.3 below, the results for the original equation (1.1) are not affected.

(2) The estimate (1.23) shows in particular that the solution $u(t)$ converges to $\alpha^*\varphi^*$ like $t\mathrm{e}^{-t/4}$ as $t \to +\infty$. As was already mentioned, the decay rate $\mathrm{e}^{-t/4}$ corresponds to the spectral gap of the linear operator $\mathcal{L}_\infty$ in $\mathrm{H}^1(\mathbf{R}_+, (1+x)^6\,\mathrm{d}x)$ and is thus optimal in our function space. The same argument suggests that this rate could be improved up to $\mathrm{e}^{-t/2}$ at the expense of assuming a faster decay of u, v as $x \to +\infty$, as in [Ga].

(3) Theorem 1.2 does not give a satisfactory estimate of the quantity $v(t) - \alpha^*\psi^*$ in L^2_t. If ε is sufficiently small, arguing as in section 3 and using three additional pairs of functionals, one can show that $\int_0^\infty (x+x^2)|v(x,t) - \alpha^*\psi^*(x)|^2\,\mathrm{d}x$ decays at least like $(1+t)^2\mathrm{e}^{-t/2}$ and that the expression

$$\int_{-\infty}^0 \mathrm{e}^{2\kappa x \mathrm{e}^{t/2}}|v(x,t) - \alpha^*\psi^*(x)|^2\,\mathrm{d}x + \int_0^\infty |v(x,t) - \alpha^*\psi^*(x)|^2\,\mathrm{d}x$$

is bounded by a polynomial in t. Since these estimates are probably not optimal and were obtained for small ε only, the corresponding calculations will not be given here.

(4) Given $\varepsilon_0 > 0$ and a nonlinearity $\mathcal{F}$ satisfying (1.2), it is straightforward to verify that all the statements in that which follows (and their proofs) hold uniformly in ε for $\varepsilon \in (0, \varepsilon_0]$. In particular, the constants t_0, δ_0, C appearing in Theorem 1.2 are independent of ε for $\varepsilon \in (0, \varepsilon_0]$. As a consequence, taking the limit $\varepsilon \to 0$ in (1.23), we obtain a local stability result for the traveling waves of the parabolic equation

(1.1) with $\varepsilon = 0$. Except for the use of slightly different function spaces, this result coincides with Theorem 1.1 in [Ga].

Combining Theorem 1.2 and Lemma 2.4 below, we obtain in particular the following convergence result for the perturbation in the original variables.

COROLLARY 1.3. *Assume that the nonlinearity $\mathcal{F}$ satisfies* (1.2)*, and let $\varepsilon > 0$. Then there exists $\delta_1 > 0$ such that, for all initial data $(w_0, w_1) \in Y$ satisfying $\|w_0\|_{Y^1}^2 + \varepsilon\|w_1\|_{Y^0}^2 \le \delta_1^2$,* (1.6) *has a unique solution $(w, w_T) \in \mathcal{C}([0,+\infty), Y)$ such that $(w(0), w_T(0)) = (w_0, w_1)$. In addition, there exists $\alpha \in \mathbf{R}$ such that*

$$\sup_{\xi\in\mathbf{R}} \left(1 + \frac{\mathrm{e}^{c_*\xi/2}}{1+|\xi|}\right) \left| w(\xi,T) - \frac{\alpha}{T^{3/2}}\, h'(\xi)\varphi^* \left(\frac{\xi\sqrt{1+\varepsilon c_*^2}}{\sqrt{T}}\right)\right| = \mathcal{O}(T^{-7/4}\log T)$$

as $T \to +\infty$.

The rest of this paper is devoted to the proof of Theorem 1.2, which is organized as follows. First, we show that the Cauchy problem for (1.16) is locally well posed in the space Z_t, in the sense of Definition 1.1. Then, in section 2.1, we decompose the solutions (u, v) of (1.16) using a spectral projection of the time-dependent operator $\mathcal{L}_t$ defined in (2.3) below. The first term in this decomposition is one-dimensional and converges to $\alpha^*(\varphi^*, \psi^*)$ as $t \to +\infty$. The remainder (f, g) satisfies an evolution system similar to (1.16), with additional terms which are estimated in section 2.2. The core of the proof is section 3, where the evolution of (f, g) in Z_t is controlled using a hierarchy of energy functionals. As in [GR2], some of these quantities are constructed in terms of the primitives (F, G) rather than the functions (f, g) themselves. Finally, the results are summarized in the short section 4.

Although the proof we present here is certainly not simple, we believe that our approach is a systematic and very convenient way to study the long-time asymptotics in a large class of dissipative systems. As a matter of fact, the present proof follows exactly the same lines as in [GR2], although the problems considered therein are significantly different. When compared with other accurate techniques, such as the renormalization group used in [BK1, Ga], our method shows at least two advantages. First, we do not need precise estimates of the resolvent of the linearized operator around the traveling wave (although some spectral information is used to construct our energy functionals). This substantial simplification is especially interesting in the perspective of possible applications to higher-dimensional problems, where standard tools like the Evans function are not available. Next, while most of our effort is devoted to controlling the linear terms in (1.16), the nonlinearities are naturally incorporated into the scheme and do not require any extra argument. In the present case, the factor $\mathrm{e}^{-t/2}$ in front of the last term in (1.16) clearly shows that the nonlinearity is irrelevant for the long-time behavior, provided the solution $u(t)$ stays globally bounded. On the other hand, a minor drawback of our approach is the introduction of nonautonomous systems and time-dependent function spaces through the change of variables (1.14). We shall avoid this difficulty by returning to the original variables to show that the Cauchy problem for (1.16) is locally well posed and to prove that our energy functionals are differentiable in time.

Notation. In that which follows, we denote by C a generic positive constant which may differ from place to place, while numbered constants $C_i, K_i, \dots$ keep the same values throughout the paper.

2. Preliminaries. We begin with a local existence result for the solutions W of (1.8) in the function space $\mathrm{Z}_0 = \mathrm{H}_0^1 \times \mathrm{L}_0^2$. We recall that H_0^1, L_0^2 are defined by the norms (1.17) with $t = 0$.

LEMMA 2.1. *Let $\varepsilon > 0$ and $\delta > 0$. There exists $\hat\tau > 0$ such that, for all initial data $(W_0, \dot W_0) \in \mathrm{Z}_0$ with $\|(W_0, \dot W_0)\|_{\mathrm{Z}_0} \le \delta$, (1.8) has a unique (mild) solution $W \in \mathcal{C}([0,\hat\tau], \mathrm{H}^1_0) \cap \mathcal{C}^1([0,\hat\tau], \mathrm{L}^2_0)$ satisfying $(W(0), W_\tau(0)) = (W_0, \dot W_0)$. The solution (W, W_τ) depends continuously on the initial data in Z_0, uniformly in $\tau \in [0,\hat\tau]$. In addition, if $(W_0, \dot W_0) \in \mathrm{H}^2_0 \times \mathrm{H}^1_0$, then $W \in \mathcal{C}([0,\hat\tau], \mathrm{H}^2_0) \cap \mathcal{C}^1([0,\hat\tau], \mathrm{H}^1_0) \cap \mathcal{C}^2([0,\hat\tau], \mathrm{L}^2_0)$ is a classical solution of (1.8) in L^2_0.*

Proof. Let $q \in \mathcal{C}^\infty(\mathbf{R})$ be a positive function satisfying $q(\xi) = \mathrm{e}^{-\kappa\xi}$ for $\xi \le 0$ and $q(\xi) = \xi^{-3}$ for $\xi \ge 1$. Setting $W(\xi,\tau) = q(\xi)\omega(\xi,\tau)$ in (1.8), we obtain for ω the equation

$$\eta\omega_{\tau\tau} - 2\nu\omega_{\xi\tau} \,=\, \omega_{\xi\xi} + \mathcal{M}(\omega, \omega_\xi, \omega_\tau)\ , \tag{2.1}$$

where

$$\begin{aligned}\mathcal{M}(\omega, \omega_\xi, \omega_\tau) \,=\, &-\left(1 - \nu\gamma - 2\nu\frac{q'}{q}\right)\omega_\tau + \left(\gamma + \frac{2q'}{q}\right)\omega_\xi + \left(\gamma\frac{q'}{q} + \frac{q''}{q}\right)\omega \\ &+\, h'q\omega^2\mathcal{N}(h, h'q\omega)\ .\end{aligned} \tag{2.2}$$

Since the functions γ, q'/q, q''/q, and $h'q$ are all bounded, and since the nonlinearity $\mathcal{F}$ in (1.1) is $\mathcal{C}^2$, it is straightforward to verify that the map $\mathcal{M} : \mathrm{H}^1(\mathbf{R}) \times \mathrm{L}^2(\mathbf{R}) \to \mathrm{L}^2(\mathbf{R})$ defined by $(\omega, \omega_\tau) \mapsto \mathcal{M}(\omega, \omega_\xi, \omega_\tau)$ is locally Lipschitz, uniformly on bounded subsets. Therefore, by a classical result [CH], the Cauchy problem for (2.1) is locally well posed in $\mathrm{H}^1 \times \mathrm{L}^2$. More precisely, for any $r > 0$, there exists $\hat\tau > 0$ such that, for all initial data $(\omega_0, \dot\omega_0) \in \mathrm{H}^1 \times \mathrm{L}^2$ with $\|(\omega_0, \dot\omega_0)\|_{\mathrm{H}^1\times\mathrm{L}^2} \le r$, (2.1) has a unique (mild) solution $\omega \in \mathcal{C}([0,\hat\tau], \mathrm{H}^1) \cap \mathcal{C}^1([0,\hat\tau], \mathrm{L}^2)$ satisfying $(\omega(0), \omega_\tau(0)) = (\omega_0, \dot\omega_0)$. This solution depends continuously on the initial data in $\mathrm{H}^1 \times \mathrm{L}^2$, uniformly in $\tau \in [0,\hat\tau]$. Moreover, if $(\omega_0, \dot\omega_0) \in \mathrm{H}^2 \times \mathrm{H}^1$, then $\omega \in \mathcal{C}([0,\hat\tau], \mathrm{H}^2) \cap \mathcal{C}^1([0,\hat\tau], \mathrm{H}^1) \cap \mathcal{C}^2([0,\hat\tau], \mathrm{L}^2)$ is a classical solution of (2.1). Thus, returning to the original function $W = q\omega$ and using the fact that

$$C^{-1}\|(\omega, \omega_\tau)\|_{\mathrm{H}^1\times\mathrm{L}^2} \,\le\, \|(W, W_\tau)\|_{\mathrm{Z}_0} \,\le\, C\|(\omega, \omega_\tau)\|_{\mathrm{H}^1\times\mathrm{L}^2}$$

for some $C \ge 1$, we obtain the desired result, if $r = C\delta$. This concludes the proof of Lemma 2.1. □

As a consequence of Definition 1.1 and Lemma 2.1, we obtain the following existence result for the solution (u, v) of (1.16).

PROPOSITION 2.2. *Let $\varepsilon > 0$, $\delta_1 > 0$, $t_2 > t_0$. There exists $T > 0$ such that, for all $t_1 \in [t_0, t_2]$ and all $(u_1, v_1) \in \mathrm{Z}_{t_1}$ satisfying $\Phi_\eta(t_1, u_1, v_1) \le \delta_1^2$, the system (1.16) has a unique solution $(u, v) \in \mathcal{C}([t_1, t_1+T], \mathrm{Z}_t)$ with initial data $(u(t_1), v(t_1)) = (u_1, v_1)$.*

Remark. In particular, Proposition 2.2 implies that, if $(u, v) \in \mathcal{C}([t_0, t_*), \mathrm{Z}_t)$ is a maximal solution of (1.16) and if $\Phi_\eta(t, u(t), v(t)) \le \delta_1^2$ for all $t \in [t_0, t_*)$, then actually $t_* = +\infty$; i.e, the solution (u, v) is globally defined.

Proof. Given $t_1 \in [t_0, t_2]$ and $(u_1, v_1) \in \mathrm{Z}_{t_1}$ satisfying $\Phi_\eta(t_1, u_1, v_1) \le \delta_1^2$, we define

$$W_1(\xi) \,=\, \mathrm{e}^{-3t_1/2}u_1(\xi\mathrm{e}^{-t_1/2})\ , \quad \dot W_1(\xi) \,=\, \mathrm{e}^{-5t_1/2}v_1(\xi\mathrm{e}^{-t_1/2})\ , \quad \xi \in \mathbf{R}\ .$$

Then $(W_1, \dot W_1) \in \mathrm{Z}_0$, and there exists a constant $C > 0$ (depending on η and t_2) such that $\|(W_1, \dot W_1)\|_{\mathrm{Z}_0} \le C\delta_1$. Since (1.8) is autonomous, it follows from Lemma 2.1 that there exists a time $\hat\tau > 0$, depending on η, $C\delta_1$ but not on $(W_1, \dot W_1)$, such that (1.8)

has a unique (mild) solution $W \in \mathcal{C}([\mathrm{e}^{t_1}, \mathrm{e}^{t_1}+\hat\tau], \mathrm{H}^1_0) \cap \mathcal{C}^1([\mathrm{e}^{t_1}, \mathrm{e}^{t_1}+\hat\tau], \mathrm{L}^2_0)$ satisfying $W(\xi, \mathrm{e}^{t_1}) = W_1(\xi)$, $W_\tau(\xi, \mathrm{e}^{t_1}) = \dot W_1(\xi)$. Now, we set $T = \log(1+\hat\tau \mathrm{e}^{-t_2})$, and for all $t \in [t_1, t_1+T] \subset [t_1, \log(\mathrm{e}^{t_1}+\hat\tau)]$ we define

$$u(x,t) \,=\, \mathrm{e}^{3t/2} W(x\mathrm{e}^{t/2}, \mathrm{e}^t) \;, \quad v(x,t) \,=\, \mathrm{e}^{5t/2} W_\tau(x\mathrm{e}^{t/2}, \mathrm{e}^t) \;.$$

By Definition 1.1, $(u,v) \in \mathcal{C}([t_1, t_1+T], \mathrm{Z}_1)$ is a solution of (1.16) with $(u(t_1), v(t_1)) = (u_1, v_1)$, and the uniqueness of this solution follows from the uniqueness of W as a mild solution of (1.8). This concludes the proof of Proposition 2.2. □

2.1. Spectral decomposition of the solution. From now on, we assume that $(u,v) \in \mathcal{C}([t_0, t_1], \mathrm{Z}_t)$ is a solution of (1.16) in the sense of Proposition 2.2. Inspired by [Ga] and [GR2], we shall decompose this solution using a spectral projection of the (time-dependent) linear operator

$$\mathcal{L}_t \,=\, \partial_x^2 + \left(\frac{x}{2} + \mathrm{e}^{t/2} \gamma(x\mathrm{e}^{t/2}) \right) \partial_x + \frac{3}{2} \;. \tag{2.3}$$

As was already mentioned, the system (1.16) is formally equivalent to the linear equation $u_t = \mathcal{L}_t u$ in the limit $t \to +\infty$. Remark that the function φ^* defined in (1.21) is an approximate eigenfunction of $\mathcal{L}_t$, in the sense that $\|\mathcal{L}_t \varphi^*\|_{\mathrm{L}^2_t} = \mathcal{O}(\mathrm{e}^{-t/4})$ as $t \to +\infty$. The corresponding spectral projection in L^2_t is given by the formula

$$u \mapsto \left(\int_{\mathbf{R}} \mathrm{e}^{-t} p(x\mathrm{e}^{t/2}) u(x) \, \mathrm{d}x \right) \varphi^*, \tag{2.4}$$

where $p : \mathbf{R} \to \mathbf{R}$ is the (unique) solution of the differential problem

$$p'(\xi) \,=\, \gamma(\xi) p(\xi) \;, \quad \xi \in \mathbf{R} \;, \quad \lim_{\xi \to +\infty} \frac{p(\xi)}{\xi^2} \,=\, 1 \;. \tag{2.5}$$

It follows from (1.11), (2.5) that $p(\xi) > 0$ for all $\xi \in \mathbf{R}$, and $p(\xi) = \mathcal{O}(\mathrm{e}^{\gamma_- \xi})$ as $\xi \to -\infty$.

Remark. Our choice of the projection (2.4) may be understood as follows. If u and W are related through (1.15), the equation $u_t = \mathcal{L}_t u$ ($t \geq t_0$) is equivalent to $W_\tau = W_{\xi\xi} + \gamma(\xi) W_\xi$ ($\tau \geq 0$). In [Ga, section 3], it is shown that the solution W of this linear equation satisfies

$$W(\xi,\tau) \,=\, \frac{1}{(\tau+\tau_0)^{3/2}} \, \varphi^* \left(\frac{\xi}{\sqrt{\tau+\tau_0}} \right) \int_{\mathbf{R}} p(\xi') W(\xi', 0) \, \mathrm{d}\xi' + \mathcal{O}\left(\frac{1}{\tau^2} \right) \;, \quad \tau \to +\infty \;,$$

where $\tau_0 \geq 0$ is arbitrary. Choosing $\tau_0 = \mathrm{e}^{t_0}$ and returning to the scaling variables (x,t), we obtain

$$u(x,t) \,=\, \varphi^*(x) \int_{\mathbf{R}} \mathrm{e}^{-t_0} p(y\mathrm{e}^{t_0/2}) u(y, t_0) \, \mathrm{d}y + \mathcal{O}(\mathrm{e}^{-t/2}) \;, \quad t \to +\infty \;.$$

This formula clearly shows the relevance of the projection (2.4) for the long-time asymptotics of the solutions of the linear equation $u_t = \mathcal{L}_t u$.

Motivated by (2.4), we introduce the functions

$$\varphi(x,t) \,=\, \frac{\varphi^*(x)}{1+\zeta(t)} \;, \quad \psi(x,t) \,=\, \varphi_t(x,t) - \frac{x}{2} \varphi_x(x,t) - \frac{3}{2} \varphi(x,t) \;, \tag{2.6}$$

where

$$\zeta(t) \,=\, \int_{\mathbf{R}} \mathrm{e}^{-t} p(x\mathrm{e}^{t/2})\varphi^*(x)\,\mathrm{d}x - 1 \;. \tag{2.7}$$

We shall show in the proof of Lemma 2.5 below that $\zeta(t)$ and $\zeta'(t)$ converge to zero as $t \to +\infty$, so that $\varphi(x,t) \to \varphi^*(x)$ and $\psi(x,t) \to \psi^*(x)$, where ψ^* is given by (1.22). By construction, we also have

$$\int_{\mathbf{R}} \mathrm{e}^{-t} p(x\mathrm{e}^{t/2})\varphi(x,t)\,\mathrm{d}x \,=\, 1 \;, \quad \int_{\mathbf{R}} \mathrm{e}^{-t} p(x\mathrm{e}^{t/2})\psi(x,t)\,\mathrm{d}x \,=\, 0 \;, \quad t \geq 0 \;. \tag{2.8}$$

Using these notations, we decompose the solution (u,v) of (1.16) as

$$u(x,t) \,=\, \alpha(t)\varphi(x,t) + f(x,t) \;, \quad v(x,t) \,=\, \beta(t)\varphi(x,t) + \alpha(t)\psi(x,t) + g(x,t) \;, \tag{2.9}$$

where

$$\alpha(t) \,=\, \int_{\mathbf{R}} \mathrm{e}^{-t} p(x\mathrm{e}^{t/2})u(x,t)\,\mathrm{d}x \;, \quad \beta(t) \,=\, \int_{\mathbf{R}} \mathrm{e}^{-t} p(x\mathrm{e}^{t/2})v(x,t)\,\mathrm{d}x \;. \tag{2.10}$$

In view of (2.8), (2.10), the functions f, g satisfy the "orthogonality relations"

$$\int_{\mathbf{R}} \mathrm{e}^{-t} p(x\mathrm{e}^{t/2})f(x,t)\,\mathrm{d}x \,=\, 0 \;, \quad \int_{\mathbf{R}} \mathrm{e}^{-t} p(x\mathrm{e}^{t/2})g(x,t)\,\mathrm{d}x \,=\, 0 \;. \tag{2.11}$$

We now determine the evolution equations satisfied by α, β, f, g. Our first result is as follows.

LEMMA 2.3. *If $(u,v) \in \mathcal{C}([t_0,t_1],\mathrm{Z}_t)$ is a solution of (1.16), then $\alpha \in \mathcal{C}^2([t_0,t_1])$ and*

$$\frac{\mathrm{d}}{\mathrm{d}t}\alpha(t) \,=\, \beta(t) \;, \quad \frac{\mathrm{d}}{\mathrm{d}t}\big(\eta \mathrm{e}^{-t}\beta(t) + \alpha(t)\big) \,=\, m(t) \;, \tag{2.12}$$

where

$$m(t) \,=\, \int_{\mathbf{R}} \mathrm{e}^{-t} p(x\mathrm{e}^{t/2})\big(-\nu\gamma(x\mathrm{e}^{t/2})v(x,t) + \mathrm{e}^{-t/2}h'(x\mathrm{e}^{t/2})u(x,t)^2 N(x,t)\big)\,\mathrm{d}x \;.$$

Proof. Let $\tau_1 = \mathrm{e}^{t_1} - \tau_0$, and let $W(\xi,\tau)$ be given by (1.15) for $\tau \in [0,\tau_1]$. By Definition 1.1, $W \in \mathcal{C}([0,\tau_1],\mathrm{H}^1_0) \cap \mathcal{C}^1([0,\tau_1],\mathrm{L}^2_0)$ is a (mild) solution of (1.8). Since $\alpha(t) = \int_{\mathbf{R}} p(\xi)W(\xi,\mathrm{e}^t - \tau_0)\,\mathrm{d}\xi$, it follows that $\alpha \in \mathcal{C}^1([t_0,t_1])$, and

$$\frac{\mathrm{d}}{\mathrm{d}t}\alpha(t) \,=\, \mathrm{e}^t \int_{\mathbf{R}} p(\xi)W_\tau(\xi,\mathrm{e}^t - \tau_0)\,\mathrm{d}\xi \,=\, \int_{\mathbf{R}} \mathrm{e}^{-t} p(x\mathrm{e}^{t/2})v(x,t)\,\mathrm{d}x \,=\, \beta(t) \;.$$

To prove that $\alpha \in \mathcal{C}^2([t_0,t_1])$, we first assume that $(W_0,\dot{W}_0) \equiv (W(\cdot,0),W_\tau(\cdot,0)) \in \mathrm{H}^2_0 \times \mathrm{H}^1_0$. Then, by Lemma 2.1, $W \in \mathcal{C}([0,\tau_1],\mathrm{H}^2_0) \cap \mathcal{C}^1([0,\tau_1],\mathrm{H}^1_0) \cap \mathcal{C}^2([0,\tau_1],\mathrm{L}^2_0)$ is a classical solution of (1.8), hence $\alpha \in \mathcal{C}^2([t_0,t_1])$, and

$$\frac{\mathrm{d}}{\mathrm{d}t}(\eta\mathrm{e}^{-t}\beta(t) + \alpha(t)) \,=\, \mathrm{e}^t \int_{\mathbf{R}} p(\xi)(\eta W_{\tau\tau} + W_\tau)(\xi,\mathrm{e}^t - \tau_0)\,\mathrm{d}\xi \,\stackrel{\mathrm{def}}{=}\, m(t) \;.$$

Since $p(\eta W_{\tau\tau} + W_\tau) = (pW_\xi)_\xi + 2\nu(pW_\tau)_\xi - \nu p\gamma W_\tau + ph'W^2\mathcal{N}(h, h'W)$ by (1.8), (2.5), we find

$$\begin{aligned} m(t) &= \mathrm{e}^t \int_{\mathbf{R}} p(\xi) \big(-\nu\gamma W_\tau + h'W^2\mathcal{N}(h, h'W)\big)(\xi, \mathrm{e}^t - \tau_0)\,\mathrm{d}\xi \\ &= \int_{\mathbf{R}} \mathrm{e}^{-t} p(x\mathrm{e}^{t/2})\big(-\nu\gamma(x\mathrm{e}^{t/2})v(x,t) + \mathrm{e}^{-t/2}h'(x\mathrm{e}^{t/2})u(x,t)^2 N(x,t)\big)\,\mathrm{d}x\ . \end{aligned}$$

For all $t \in [t_0, t_1]$, we thus have

$$\eta\mathrm{e}^{-t}\beta(t) + \alpha(t) = \eta\mathrm{e}^{-t_0}\beta(t_0) + \alpha(t_0) + \int_{t_0}^t m(s)\,\mathrm{d}s\ . \tag{2.13}$$

By Lemma 2.1, both sides of (2.13) are continuous functions of the initial data $(W_0, \dot W_0)$ in Z_0. Since (2.13) is satisfied for all $(W_0, \dot W_0)$ in the dense subspace $\mathrm{H}^2_0 \times \mathrm{H}^1_0$, the equality must hold for all $(W_0, \dot W_0) \in \mathrm{Z}_0$. This shows that $\eta\mathrm{e}^{-t}\beta + \alpha \in \mathcal{C}^1([t_0, t_1])$ and that (2.12) holds. The proof of Lemma 2.3 is complete. □

It follows from (1.16), (2.9), and Lemma 2.3 that $(f, g) \in \mathcal{C}([t_0, t_1], \mathrm{Z}_t)$ is a solution (in the sense of Definition 1.1) of the system

$$\begin{aligned} &f_t - \frac{x}{2}f_x - \frac{3}{2}f = g\ , \\ &\eta\mathrm{e}^{-t}\left(g_t - \frac{x}{2}g_x - \frac{5}{2}g\right) + (1 - \nu\gamma(x\mathrm{e}^{t/2}))g - 2\nu\mathrm{e}^{-t/2}g_x \\ &\quad = f_{xx} + \mathrm{e}^{t/2}\gamma(x\mathrm{e}^{t/2})f_x + r(x,t)\ , \end{aligned} \tag{2.14}$$

where

$$\begin{aligned} r(x,t) &= \alpha(\varphi_{xx} + \mathrm{e}^{t/2}\gamma(x\mathrm{e}^{t/2})\varphi_x - \psi) - \eta\mathrm{e}^{-t}(2\beta\psi + \alpha(\psi_t - \tfrac{x}{2}\psi_x - \tfrac{5}{2}\psi)) \\ &\quad + \nu\gamma(x\mathrm{e}^{t/2})(\beta\varphi + \alpha\psi) + 2\nu\mathrm{e}^{-t/2}(\beta\varphi_x + \alpha\psi_x) \\ &\quad + \mathrm{e}^{-t/2}h'(x\mathrm{e}^{t/2})u(x,t)^2 N(x,t) - m(t)\varphi\ . \end{aligned} \tag{2.15}$$

Using (2.5), (2.6), (2.8), and the definition of $m(t)$ in Lemma 2.3, it is not difficult to verify that

$$\int_{\mathbf{R}} \mathrm{e}^{-t} p(x\mathrm{e}^{t/2})\big(r(x,t) - \nu\gamma(x\mathrm{e}^{t/2})g(x,t)\big)\,\mathrm{d}x = 0\ . \tag{2.16}$$

Finally, as in [GR2], it will be useful to consider also the primitives

$$F(x,t) = \int_{-\infty}^x \mathrm{e}^{-t} p(y\mathrm{e}^{t/2}) f(y,t)\,\mathrm{d}y\ , \quad G(x,t) = \int_{-\infty}^x \mathrm{e}^{-t} p(y\mathrm{e}^{t/2}) g(y,t)\,\mathrm{d}y\ . \tag{2.17}$$

Using (2.11) and standard inequalities (see Lemma 2.7 below and the remark at the end of this section), it is straightforward to verify that $(F, G) \in \mathcal{C}^1([t_0, t_1], \mathrm{H}^1 \times \mathrm{L}^2)$ is a classical solution of the system

$$\begin{aligned} &F_t - \frac{x}{2}F_x = G\ , \\ &\eta\mathrm{e}^{-t}\Big(G_t - \frac{x}{2}G_x - G\Big) + G - 2\nu\mathrm{e}^{-t/2}G_x = F_{xx} - \mathrm{e}^{t/2}\gamma(x\mathrm{e}^{t/2})F_x + R(x,t)\ , \end{aligned} \tag{2.18}$$

where

$$R(x,t) = \int_{-\infty}^x \mathrm{e}^{-t} p(y\mathrm{e}^{t/2})\big(r(y,t) - \nu\gamma(y\mathrm{e}^{t/2})g(y,t)\big)\,\mathrm{d}y\ . \tag{2.19}$$

2.2. Bounds on the nonlinear terms. In this subsection, we assume that $(u,v) \in \mathcal{C}([t_0,t_1],\mathrm{Z}_t)$ is a solution of (1.16) satisfying the bound

$$\|u(t)\|_{\mathrm{H}^1_t} \,\leq\, 1\;, \quad t \in [t_0,t_1]\;. \tag{2.20}$$

Then $u(t)$ is uniformly bounded in a weighted L^∞ space, as a consequence of the following result.

LEMMA 2.4. *There exists a constant $K_0 > 0$ such that, for all $t \geq 0$ and all $w \in \mathrm{H}^1_t$,*

$$\sup_{x\leq 0} \mathrm{e}^{\kappa x \mathrm{e}^{t/2}} |w(x)| + \sup_{x\geq 0} (1{+}x)^3 |w(x)| \,\leq\, K_0 \|w\|_{\mathrm{H}^1_t}\;. \tag{2.21}$$

Remark. Note the crucial fact that the constant K_0 in (2.21) is independent of t.

Proof. Let $t \geq 0$ and $w \in \mathrm{H}^1_t$. By a classical inequality, there exists $C > 0$ such that

$$\sup_{x\geq 0} (1{+}x)^6 w(x)^2 \,\leq\, C \int_0^\infty (1{+}x)^6 (w(x)^2 + w'(x)^2)\,\mathrm{d}x\;. \tag{2.22}$$

In particular, $w(0)^2 \leq C\|w\|^2_{\mathrm{H}^1_t}$. On the other hand, we have for all $x < 0$

$$\begin{aligned} \mathrm{e}^{2\kappa x \mathrm{e}^{t/2}} w(x)^2 \,&=\, w(0)^2 - \int_x^0 \mathrm{e}^{2\kappa y \mathrm{e}^{t/2}} \big(2w(y)w'(y) + 2\kappa \mathrm{e}^{t/2} w(y)^2\big)\,\mathrm{d}y \\ &\leq\, w(0)^2 + \int_{-\infty}^0 \mathrm{e}^{2\kappa y \mathrm{e}^{t/2}} (w(y)^2 + w'(y)^2)\,\mathrm{d}y\;. \end{aligned} \tag{2.23}$$

Combining (2.22), (2.23), we obtain (2.21). This concludes the proof of Lemma 2.4. □

In that which follows, it will be natural to control the solution (u,v) of (1.16) in terms of the functions α, β, f, g defined in (2.9), (2.10). The equivalence of the corresponding norms is the content of our next result.

LEMMA 2.5. *There exists a constant $K_1 \geq 1$ such that, for all $t \geq 0$ and all $(u,v) \in \mathrm{Z}_t$,*

$$\begin{aligned} K_1^{-1} \|u\|_{\mathrm{H}^1_t} \,&\leq\, |\alpha| + \|f\|_{\mathrm{H}^1_t} \,\leq\, K_1 \|u\|_{\mathrm{H}^1_t}\;, \\ K_1^{-1} \|v\|_{\mathrm{L}^2_t} \,&\leq\, |\alpha| + |\beta| + \|g\|_{\mathrm{L}^2_t} \,\leq\, K_1 (\|u\|_{\mathrm{H}^1_t} + \|v\|_{\mathrm{L}^2_t})\;, \end{aligned} \tag{2.24}$$

where α, β are defined in (2.10) *and f, g in* (2.9).

Proof. From (1.11), we know that $\gamma(\xi) \to \gamma_-$ as $\xi \to -\infty$ and $\gamma(\xi) \sim 2/(\xi{+}\xi_0)$ as $\xi \to +\infty$. Setting $\xi_1 = -\xi_0 + 2/\gamma_-$, we decompose $\gamma(\xi)$ as $\gamma_0(\xi) + \hat\gamma(\xi)$, where

$$\gamma_0(\xi) \,=\, \begin{cases} \gamma_- & \text{if } \xi < \xi_1\;, \\ 2/(\xi{+}\xi_0) & \text{if } \xi \geq \xi_1\;. \end{cases}$$

By (1.11), the remainder $\hat\gamma(\xi)$ decays exponentially as $|\xi| \to \infty$. Thus the solution of (2.5) can be represented as

$$p(\xi) = p_0(\xi) \exp\left(- \int_\xi^\infty \hat\gamma(s)\,\mathrm{d}s \right)\;, \quad p_0(\xi) = \begin{cases} (2/\gamma_-)^2 \mathrm{e}^{\gamma_-(\xi-\xi_1)} & \text{if } \xi < \xi_1, \\ (\xi{+}\xi_0)^2 & \text{if } \xi \geq \xi_1. \end{cases} \tag{2.25}$$

In particular, there exists $C_0 \geq 1$ such that

$$p(\xi) \leq C_0 \begin{cases} \mathrm{e}^{\gamma_- \xi} & \text{if } \xi < 0 \ , \\ (1+\xi)^2 & \text{if } \xi \geq 0 \ , \end{cases} \qquad p(\xi) \geq C_0^{-1} \begin{cases} \mathrm{e}^{\gamma_- \xi} & \text{if } \xi < 0 \ , \\ (1+\xi)^2 & \text{if } \xi \geq 0 \ . \end{cases} \tag{2.26}$$

Using (2.25) and remembering that $\int_0^\infty x^2 \varphi^*(x)\,\mathrm{d}x = 1$, we decompose the function $\zeta(t)$ defined in (2.7) as

$$\begin{aligned} \zeta(t) \ &= \int_{-\infty}^0 \mathrm{e}^{-t} p(x\mathrm{e}^{t/2}) \varphi^*(x)\,\mathrm{d}x + \int_0^\infty \mathrm{e}^{-t} \big(p(x\mathrm{e}^{t/2}) - p_0(x\mathrm{e}^{t/2})\big) \varphi^*(x)\,\mathrm{d}x \\ &\quad + \int_0^\infty \big(\mathrm{e}^{-t} p_0(x\mathrm{e}^{t/2}) - x^2\big) \varphi^*(x)\,\mathrm{d}x \ = \ \zeta_1(t) + \zeta_2(t) + \zeta_3(t) \ . \end{aligned}$$

Using (1.21), we remark that

$$\zeta_1(t) = \frac{\mathrm{e}^{-3t/2}}{\sqrt{4\pi}} \int_{-\infty}^0 p(\xi)\,\mathrm{d}\xi \ , \quad \zeta_2(t) \ = \ \mathrm{e}^{-3t/2} \int_0^\infty \big(p(\xi) - p_0(\xi)\big) \varphi^*(\xi \mathrm{e}^{-t/2})\,\mathrm{d}\xi \ ,$$

where $p(\xi) - p_0(\xi)$ decays exponentially to zero as $\xi \to +\infty$ due to (2.25). On the other hand, setting $\bar\xi = \max(0, \xi_1)$, we have

$$\zeta_3(t) \ = \ \mathrm{e}^{-3t/2} \int_0^{\bar\xi} \big(p_0(\xi) - (\xi+\xi_0)^2\big) \varphi^*(\xi \mathrm{e}^{-t/2})\,\mathrm{d}\xi + \int_0^\infty \big(2\xi_0 x \mathrm{e}^{-t/2} + \xi_0^2 \mathrm{e}^{-t}\big) \varphi^*(x)\,\mathrm{d}x \ .$$

It follows immediately from these expressions that

$$|\zeta(t)| + |\zeta'(t)| + |\zeta''(t)| \ \leq \ C_1 \mathrm{e}^{-t/2} \ , \quad (1+\zeta(t))^{-1} \ \leq \ C_1 \ , \quad t \geq 0 \ , \tag{2.27}$$

for some $C_1 > 0$. As a consequence, the functions $\varphi(x,t), \psi(x,t)$ defined by (2.6) satisfy the bounds

$$\|\varphi(t)\|_{\mathrm{H}^1_t} + \|\psi(t)\|_{\mathrm{L}^2_t} \ \leq \ C_2 \ , \quad t \geq 0 \ , \tag{2.28}$$

and

$$\|\varphi(t) - \varphi^*\|_{\mathrm{H}^1_t} + \|\psi(t) - \psi^*\|_{\mathrm{L}^2_t} \ \leq \ C_2 \mathrm{e}^{-t/2} \ , \quad t \geq 0 \ , \tag{2.29}$$

for some $C_2 > 0$.

Now, let $t \geq 0$, $(u,v) \in \mathrm{Z}_t$, and let α, β be defined as in (2.10). In view of (2.26), we have

$$\begin{aligned} |\alpha| \ &\leq \ C_0 \int_0^\infty (1+x)^2 |u|\,\mathrm{d}x + C_0 \mathrm{e}^{-t} \int_{-\infty}^0 \mathrm{e}^{\gamma_- x \mathrm{e}^{t/2}} |u|\,\mathrm{d}x \\ &\leq \ C_0 \left(\int_0^\infty (1+x)^6 u^2\,\mathrm{d}x \right)^{1/2} + \frac{C_0 \mathrm{e}^{-5t/4}}{(\gamma_- - \kappa)^{1/2}} \left(\int_{-\infty}^0 \mathrm{e}^{2\kappa x \mathrm{e}^{t/2}} u^2\,\mathrm{d}x \right)^{1/2} \ ; \end{aligned} \tag{2.30}$$

hence $|\alpha| \leq C_3 \|u\|_{\mathrm{L}^2_t}$ for some $C_3 > 0$. Similarly, we have $|\beta| \leq C_3 \|v\|_{\mathrm{L}^2_t}$. Using these bounds together with (2.9), (2.28), we obtain (2.24). This concludes the proof of Lemma 2.5. □

We now estimate the remainder terms $m(t)$ and $r(x,t)$ in (2.12), (2.14).

LEMMA 2.6. *There exists a constant $K_2 > 0$ such that, if $(u,v) \in \mathcal{C}([t_0,t_1], \mathrm{Z}_t)$ is a solution of* (1.16) *satisfying* (2.20)*, then*

$$\|r(t)\|_{\mathrm{L}^2_t} + \mathrm{e}^{t/4} |m(t)| \ \leq \ K_2 \mathrm{e}^{-t/4} \big(\alpha(t)^2 + \beta(t)^2 + \|f(t)\|^2_{\mathrm{H}^1_t} + \|g(t)\|^2_{\mathrm{L}^2_t}\big)^{1/2} \tag{2.31}$$

for all $t \in [t_0, t_1]$, where $r(x,t)$ is defined in (2.15) *and $m(t)$ in Lemma* 2.3.

Proof. We first consider the function $r_1(x,t) = \varphi_{xx} + \mathrm{e}^{t/2}\gamma(x\mathrm{e}^{t/2})\varphi_x - \psi$. It follows from (1.21), (2.6) that $r_1(x,t) = (1+\zeta(t))^{-1}(\hat r(x,t) + \zeta'(t)\varphi(x,t))$, where

$$\hat r(x,t) \,=\, \begin{cases} (\mathrm{e}^{t/2}\gamma(x\mathrm{e}^{t/2}) - 2/x)\varphi_x^* & \text{if } x > 0\ , \\ 3\varphi^*/2 & \text{if } x \le 0\ . \end{cases}$$

By (2.27), (2.28), we have $\|\zeta'(t)\varphi(t)\|_{\mathrm{L}^2_t} \le C_1C_2\mathrm{e}^{-t/2}$. To bound $\hat r(x,t)$, we observe that the function $\xi \mapsto (2-\xi\gamma(\xi))$ belongs to $\mathrm{L}^2(\mathbf{R}_+)$ by (1.11). Since $\varphi_x^* = -(x/2)\varphi^*$ for $x > 0$, we thus find

$$\int_0^\infty (1+x)^6 \hat r(x,t)^2\,\mathrm{d}x \,\le\, \frac{\mathrm{e}^{-t/2}}{4}\left(\sup_{x\ge 0}(1+x)^6\varphi^*(x)^2\right)\int_0^\infty (2-\xi\gamma(\xi))^2\,\mathrm{d}\xi\ ,$$
$$\int_{-\infty}^0 \mathrm{e}^{2\kappa x\mathrm{e}^{t/2}}\hat r(x,t)^2\,\mathrm{d}x = \frac{9}{32\pi\kappa}\,\mathrm{e}^{-t/2}\ .$$

Summarizing, we obtain $\|r_1(t)\|_{\mathrm{L}^2_t} \le C_4\mathrm{e}^{-t/4}$ for some $C_4 > 0$. Similarly, since $\gamma \in \mathrm{L}^2(\mathbf{R}_+) \cap \mathrm{L}^\infty(\mathbf{R}_-)$, we find $\|\gamma(x\mathrm{e}^{t/2})\varphi(t)\|_{\mathrm{L}^2_t} \le C_4\mathrm{e}^{-t/4}$ and $\|\gamma(x\mathrm{e}^{t/2})\psi(t)\|_{\mathrm{L}^2_t} \le C_4\mathrm{e}^{-t/4}$.

We next bound the nonlinear term $r_2(x,t) = \mathrm{e}^{-t/2}h'(x\mathrm{e}^{t/2})u(x,t)^2N(x,t)$, where $N(x,t) = \mathcal{N}(h(x\mathrm{e}^{t/2}), \mathrm{e}^{-3t/2}h'(x\mathrm{e}^{t/2})u(x,t))$. In view of (1.10), (2.20), (2.21), there exists $C_5 > 0$ such that $\sup_{x\in\mathbf{R}}|h'(x\mathrm{e}^{t/2})u(x,t)| \le C_5$ for all $t \in [t_0,t_1]$. In particular, since $\mathcal{N} : \mathbf{R}^2 \to \mathbf{R}$ is continuous, we have $\|N(\cdot,t)\|_{\mathrm{L}^\infty} \le N_0$ for some $N_0 > 0$ and all $t \in [t_0,t_1]$. It follows that $\|r_2\|_{\mathrm{L}^2_t} \le \mathrm{e}^{-t/2}C_5N_0\|u(t)\|_{\mathrm{L}^2_t}$ for $t \in [t_0,t_1]$.

Finally, the function $m(t)$ defined in Lemma 2.3 can be written as $m_1(t) + m_2(t)$, where

$$m_1(t) \,=\, -\nu\int_{\mathbf{R}} \mathrm{e}^{-t}p(x\mathrm{e}^{t/2})\gamma(x\mathrm{e}^{t/2})v(x,t)\,\mathrm{d}x\ , \quad m_2(t) \,=\, \int_{\mathbf{R}} \mathrm{e}^{-t}p(x\mathrm{e}^{t/2})r_2(x,t)\,\mathrm{d}x\ .$$

Proceeding as in (2.30), we find that $|m_2(t)| \le C_3\|r_2(t)\|_{\mathrm{L}^2_t} \le \mathrm{e}^{-t/2}C_3C_5N_0\|u(t)\|_{\mathrm{L}^2_t}$. Moreover, since $\mathrm{e}^{-t}\gamma(x\mathrm{e}^{t/2})p(x\mathrm{e}^{t/2}) \le C\mathrm{e}^{-t/2}(1+x)$ for $x \ge 0$, we obtain

$$|m_1(t)| \,\le\, C\nu\mathrm{e}^{-t/2}\int_0^\infty (1+x)|v(x,t)|\,\mathrm{d}x + C\nu\mathrm{e}^{-t}\int_{-\infty}^0 \mathrm{e}^{\gamma_- x\mathrm{e}^{t/2}}|v(x,t)|\,\mathrm{d}x\ ;$$

hence $|m_1(t)| \le C_6\nu\mathrm{e}^{-t/2}\|v(t)\|_{\mathrm{L}^2_t}$ for some $C_6 > 0$. Therefore, there exists $C_7 > 0$ such that

$$|m(t)| \le C_7\mathrm{e}^{-t/2}(\|u(t)\|_{\mathrm{L}^2_t} + \|v(t)\|_{\mathrm{L}^2_t})\ , \quad t \in [t_0,t_1]\ . \tag{2.32}$$

Summarizing our results and observing that the functions φ, φ_x, ψ, ψ_x, ψ_t, $x\psi_x$ are uniformly bounded in L^2_t by (2.6), (2.27), we see that the remainder $r(x,t)$ defined by (2.15) satisfies

$$\|r(t)\|_{\mathrm{L}^2_t} \,\le\, C_8\mathrm{e}^{-t/4}\big(|\alpha(t)| + |\beta(t)| + \|u(t)\|_{\mathrm{H}^1_t} + \|v(t)\|_{\mathrm{L}^2_t}\big)\ , \quad t \in [t_0,t_1]\ , \tag{2.33}$$

for some $C_8 > 0$. Combining (2.24), (2.32), (2.33), we obtain (2.31). This concludes the proof of Lemma 2.6. □

Finally, we bound the primitives F, G, R defined in (2.17), (2.19).

LEMMA 2.7. *There exists a constant $K_3 > 0$ such that, for all $t \geq 0$ and all $f \in \mathrm{L}^2_t$ satisfying $\int_{\mathbf{R}} p(x\mathrm{e}^{t/2})f(x)\,\mathrm{d}x = 0$, the following estimate holds:*

$$\int_{\mathbf{R}} \left(1 + \frac{\mathrm{e}^t}{p(x\mathrm{e}^{t/2})}\right) F^2\,\mathrm{d}x \leq K_3 \left(\mathrm{e}^{-2t}\int_{-\infty}^0 \mathrm{e}^{2\kappa x\mathrm{e}^{t/2}} f^2\,\mathrm{d}x + \int_0^\infty (1+x)^6 f^2\,\mathrm{d}x\right), \tag{2.34}$$

where $F(x) = \int_{-\infty}^x \mathrm{e}^{-t} p(y\mathrm{e}^{t/2})f(y)\,\mathrm{d}y$.

Proof. Let $t \geq 0$ and $f \in \mathrm{L}^2_t$. We start from the identity

$$\mathrm{e}^t \int_{-\infty}^0 \mathrm{e}^{-\gamma_- x\mathrm{e}^{t/2}} F(x)^2\,\mathrm{d}x + \frac{\mathrm{e}^{t/2}}{\gamma_-}F(0)^2 \;=\; \frac{2\mathrm{e}^{-t/2}}{\gamma_-}\int_{-\infty}^0 \mathrm{e}^{-\gamma_- x\mathrm{e}^{t/2}} p(x\mathrm{e}^{t/2})F(x)f(x)\,\mathrm{d}x\;,$$

which is a simple integration by parts. Applying Hölder's inequality to the right-hand side, we obtain

$$\mathrm{e}^t \int_{-\infty}^0 \mathrm{e}^{-\gamma_- x\mathrm{e}^{t/2}} F(x)^2\,\mathrm{d}x + \frac{\mathrm{e}^{t/2}}{\gamma_-}F(0)^2 \;\leq\; \frac{4\mathrm{e}^{-2t}}{\gamma_-^2}\int_{-\infty}^0 \mathrm{e}^{-\gamma_- x\mathrm{e}^{t/2}} p(x\mathrm{e}^{t/2})^2 f(x)^2\,\mathrm{d}x\;.$$

Using (2.26) and remembering that $\gamma_- = c + 2\kappa > 2\kappa$, we conclude that

$$\int_{-\infty}^0 \left(1 + \frac{\mathrm{e}^t}{p(x\mathrm{e}^{t/2})}\right) F(x)^2\,\mathrm{d}x + \mathrm{e}^{t/2}F(0)^2 \;\leq\; C\mathrm{e}^{-2t}\int_{-\infty}^0 \mathrm{e}^{2\kappa x\mathrm{e}^{t/2}} f(x)^2\,\mathrm{d}x \tag{2.35}$$

for some $C > 0$.

Since $\int_{\mathbf{R}} p(x\mathrm{e}^{t/2})f(x)\,\mathrm{d}x = 0$, we have $F(x) = -\int_x^\infty \mathrm{e}^{-t} p(y\mathrm{e}^{t/2})f(y)\,\mathrm{d}y$. Using (2.26) and a classical inequality of Hardy [HLP, Theorem 328], we find

$$\int_0^\infty F(x)^2\,\mathrm{d}x \;\leq\; 4\int_0^\infty \mathrm{e}^{-2t}x^2 p(x\mathrm{e}^{t/2})^2 f(x)^2\,\mathrm{d}x \;\leq\; 4C_0^2\int_0^\infty (1+x)^6 f(x)^2\,\mathrm{d}x\;. \tag{2.36}$$

On the other hand, since $F(x) = F(0) + \int_0^x \mathrm{e}^{-t} p(y\mathrm{e}^{t/2})f(y)\,\mathrm{d}y$, we have for $x > 0$

$$\frac{\mathrm{e}^{t/2}|F(x)|}{1 + x\mathrm{e}^{t/2}} \;\leq\; \frac{\mathrm{e}^{t/2}|F(0)|}{1 + x\mathrm{e}^{t/2}} + \frac{C_0}{x}\int_0^x (1+y)^2|f(y)|\,\mathrm{d}y\;. \tag{2.37}$$

Using another form of Hardy's inequality [HLP, Theorem 327], we thus obtain

$$\int_0^\infty \frac{\mathrm{e}^t F(x)^2}{(1+x\mathrm{e}^{t/2})^2}\,\mathrm{d}x \;\leq\; 2\mathrm{e}^{t/2}F(0)^2 + 8C_0^2\int_0^\infty (1+x)^4 f(x)^2\,\mathrm{d}x\;. \tag{2.38}$$

Combining (2.35), (2.36), (2.38), and using (2.26), we arrive at (2.34). This concludes the proof of Lemma 2.7. □

LEMMA 2.8. *There exists a constant $K_4 > 0$ such that, if $(u, v) \in \mathcal{C}([t_0, t_1], \mathrm{Z}_t)$ is a solution of (1.16) satisfying (2.20), then*

$$\int_{\mathbf{R}} \left(1 + \frac{\mathrm{e}^t}{p(x\mathrm{e}^{t/2})}\right) R^2\,\mathrm{d}x \;\leq\; K_4\mathrm{e}^{-t/2}\big(\alpha(t)^2 + \beta(t)^2 + \|f(t)\|^2_{\mathrm{H}^1_t} + \|g(t)\|^2_{\mathrm{L}^2_t}\big) \tag{2.39}$$

for all $t \in [t_0, t_1]$, where $R(x,t)$ is defined in (2.19).

Proof. Following the proof of Lemma 2.7, we obtain as in (2.35)

$$\int_{-\infty}^0 \left(1 + \frac{\mathrm{e}^t}{p(x\mathrm{e}^{t/2})}\right) R(x,t)^2\,\mathrm{d}x + \mathrm{e}^{t/2}R(0,t)^2 \;\leq\; C\mathrm{e}^{-2t}\int_{-\infty}^0 \mathrm{e}^{2\kappa x\mathrm{e}^{t/2}}(r^2 + \nu^2\gamma_-^2 g^2)\,\mathrm{d}x\;.$$

Next, remarking that $\mathrm{e}^{-t}p(x\mathrm{e}^{t/2})\gamma(x\mathrm{e}^{t/2}) \le C\mathrm{e}^{-t/2}(1+x)$ for $x \ge 0$, we find instead of (2.36), (2.38)

$$\int_0^\infty R(x,t)^2\,\mathrm{d}x \,\le\, C\int_0^\infty \big((1+x)^6 r(x,t)^2 + \nu^2\mathrm{e}^{-t}(1+x)^4 g(x,t)^2\big)\,\mathrm{d}x\ ,$$
$$\int_0^\infty \frac{\mathrm{e}^t R(x,t)^2}{(1+x\mathrm{e}^{t/2})^2}\,\mathrm{d}x \,\le\, 2\mathrm{e}^{t/2}R(0,t)^2 + C\int_0^\infty \big((1+x)^4 r^2 + \nu^2\mathrm{e}^{-t}(1+x)^2 g^2\big)\,\mathrm{d}x\ .$$

Combining these estimates, we obtain

$$\int_{\mathbf{R}} \left(1 + \frac{\mathrm{e}^t}{p(x\mathrm{e}^{t/2})}\right) R(x,t)^2\,\mathrm{d}x \,\le\, C\big(\|r(t)\|_{\mathrm{L}^2_t}^2 + \nu^2\mathrm{e}^{-t}\|g(t)\|_{\mathrm{L}^2_t}^2\big)\ ,$$

and (2.39) follows using Lemma 2.6. This concludes the proof of Lemma 2.8. □

Remark. For $t \ge 0$, $k \in \mathbf{N}$, let X^k_t be the weighted Sobolev space defined by the norm

$$\|u\|^2_{\mathrm{X}^0_t} \,=\, \int_{-\infty}^0 \mathrm{e}^{-\gamma_- x\mathrm{e}^{t/2}} u(x)^2\,\mathrm{d}x + \int_0^\infty u(x)^2\,\mathrm{d}x\ ,\quad \|u\|^2_{\mathrm{X}^k_t} \,=\, \sum_{i=0}^k \|\partial_x^i u\|^2_{\mathrm{X}^0_t}\ .$$

If $(u,v) \in \mathcal{C}([t_0,t_1],\mathrm{Z}_t)$ is a solution of (1.16), it follows from Lemma 2.7 and from the definition (2.17) of F,G that $(F,G) \in \mathrm{X}^2_t \times \mathrm{X}^1_t$ for all $t \in [t_0,t_1]$. Moreover, using a density argument as in the proof of Lemma 2.3, one can verify that $(F,G) \in \mathcal{C}^1([t_0,t_1],\mathrm{X}^1_t \times \mathrm{X}^0_t)$ is a classical solution of (2.18). As in Definition 1.1, this means that if

$$\tilde F(\xi,t) \,=\, F(\xi\mathrm{e}^{-t/2},t)\ ,\quad \tilde G(\xi,t) \,=\, G(\xi\mathrm{e}^{-t/2},t)\ ,$$

then $(\tilde F,\tilde G) \in \mathcal{C}^1([t_0,t_1],\mathrm{X}^1_0 \times \mathrm{X}^0_0) \cap \mathcal{C}([t_0,t_1],\mathrm{X}^2_0 \times \mathrm{X}^1_0)$. For later use, we also note that

$$\tilde F_t(\xi,t) \,=\, \Big(F_t - \frac{x}{2}F_x\Big)(\xi\mathrm{e}^{-t/2},t)\ ,\quad \tilde G_t(\xi,t) \,=\, \Big(G_t - \frac{x}{2}G_x\Big)(\xi\mathrm{e}^{-t/2},t)\ . \tag{2.40}$$

3. Energy estimates. As in the previous section, we assume that $(u,v) \in \mathcal{C}([t_0,t_1],\mathrm{Z}_t)$ is a solution of (1.16) satisfying the bound (2.20). To control the time behavior of the functions f,g defined in (2.9), we shall use five pairs of energy functionals. The construction of these functionals follows essentially the same lines as in [GR2], with additional complications due to the drift $\gamma(\xi)$ and to the nontrivial spectral projection (2.4).

We first introduce unweighted functionals for the primitives F,G defined in (2.18):

$$E_0(t) \,=\, \int_{\mathbf{R}} \left(\frac{1}{2}F^2 + \eta\mathrm{e}^{-t}FG\right)\mathrm{d}x\ ,\quad \mathcal{E}_0(t) \,=\, \frac{1}{2}\int_{\mathbf{R}} \big(F_x^2 + \eta\mathrm{e}^{-t}G^2\big)\,\mathrm{d}x\ . \tag{3.1}$$

Lemma 3.1. *Assume that $(u,v) \in \mathcal{C}([t_0,t_1],\mathrm{Z}_t)$ is a solution of (1.16). Then E_0 and $\mathcal{E}_0$ belong to $\mathcal{C}^1([t_0,t_1])$ and*

$$\dot E_0 \,=\, -\frac{E_0}{2} + \int_{\mathbf{R}} \left(-F_x^2 + \frac{\mathrm{e}^t}{2}\gamma'(x\mathrm{e}^{t/2})F^2 + \eta\mathrm{e}^{-t}G^2 - 2\nu\mathrm{e}^{-t/2}F_xG + FR\right)\mathrm{d}x\ ,$$
$$\dot{\mathcal{E}}_0 \,=\, \frac{\mathcal{E}_0}{2} + \int_{\mathbf{R}} \big(-G^2 - \mathrm{e}^{t/2}\gamma(x\mathrm{e}^{t/2})F_xG + GR\big)\,\mathrm{d}x$$

for all $t \in [t_0, t_1]$, where R is defined in (2.19).

Remark. Here and afterward, we use the notation $\dot{E} = (\mathrm{d}E/\mathrm{d}t)$, $\dot{\mathcal{E}} = (\mathrm{d}\mathcal{E}/\mathrm{d}t)$.

Proof. Since $(F, G) \in \mathcal{C}^1([t_0, t_1], \mathrm{H}^1 \times \mathrm{L}^2)$, the functions E_0 and $\mathcal{E}_0$ belong to $\mathcal{C}^1([t_0, t_1])$, and a direct calculation yields

$$\begin{aligned} \dot{E}_0(t) &= \int_{\mathbf{R}} \big(FF_t + \eta \mathrm{e}^{-t}((FG)_t - FG) \big) \, \mathrm{d}x \ , \\ \dot{\mathcal{E}}_0(t) &= \int_{\mathbf{R}} \big(-F_{xx}F_t + \eta \mathrm{e}^{-t}(GG_t - \tfrac{1}{2}G^2) \big) \, \mathrm{d}x \ . \end{aligned}$$

Using the identities

$$\begin{aligned} &FF_t + \eta \mathrm{e}^{-t}\big((FG)_t - \tfrac{x}{2}(FG)_x - FG - G^2\big) \\ &\qquad = FF_{xx} + (\tfrac{x}{2} - \mathrm{e}^{t/2}\gamma(x\mathrm{e}^{t/2}))FF_x + 2\nu \mathrm{e}^{-t/2} FG_x + FR \ , \end{aligned} \tag{3.2}$$

$$\begin{aligned} &-F_{xx}F_t + \eta \mathrm{e}^{-t}(GG_t - \tfrac{x}{2}GG_x - G^2) \\ &\qquad = -G^2 - \tfrac{x}{2}F_xF_{xx} - \mathrm{e}^{t/2}\gamma(x\mathrm{e}^{t/2})GF_x + 2\nu \mathrm{e}^{-t/2} GG_x + GR \ , \end{aligned} \tag{3.3}$$

which follow from (2.18), and integrating by parts, we obtain the desired expressions. This concludes the proof of Lemma 3.1. □

We next introduce weighted functionals for the primitives F, G:

$$\begin{aligned} E_1(t) &= \int_{\mathbf{R}} \frac{\mathrm{e}^t}{p(x\mathrm{e}^{t/2})} \left(\frac{1}{2}F^2 + \eta \mathrm{e}^{-t} FG \right) \mathrm{d}x \ , \\ \mathcal{E}_1(t) &= \frac{1}{2} \int_{\mathbf{R}} \frac{\mathrm{e}^t}{p(x\mathrm{e}^{t/2})} \left(F_x^2 + \eta \mathrm{e}^{-t} G^2 \right) \mathrm{d}x \ , \end{aligned} \tag{3.4}$$

where the weight p is defined in (2.5).

LEMMA 3.2. *Assume that $(u, v) \in \mathcal{C}([t_0, t_1], \mathrm{Z}_t)$ is a solution of* (1.16). *Then E_1 and $\mathcal{E}_1$ belong to $\mathcal{C}^1([t_0, t_1])$ and*

$$\begin{aligned} \dot{E}_1 &= \frac{E_1}{2} + \int_{\mathbf{R}} \frac{\mathrm{e}^t}{p(x\mathrm{e}^{t/2})} (-F_x^2 + \eta \mathrm{e}^{-t} G^2 + 2\nu \mathrm{e}^{-t/2} FG_x + FR) \, \mathrm{d}x \ , \\ \dot{\mathcal{E}}_1 &= \frac{3\mathcal{E}_1}{2} + \int_{\mathbf{R}} \frac{\mathrm{e}^t}{p(x\mathrm{e}^{t/2})} (-G^2 + 2\nu \mathrm{e}^{-t/2} GG_x + GR) \, \mathrm{d}x \end{aligned}$$

for all $t \in [t_0, t_1]$.

Proof. We remark that

$$E_1(t) = \int_{\mathbf{R}} \frac{\mathrm{e}^{t/2}}{p(\xi)} \left(\frac{1}{2}\tilde{F}^2 + \eta \mathrm{e}^{-t} \tilde{F}\tilde{G} \right) \mathrm{d}\xi \ , \quad \mathcal{E}_1(t) = \frac{1}{2} \int_{\mathbf{R}} \frac{\mathrm{e}^{3t/2}}{p(\xi)} (\tilde{F}_\xi^2 + \eta \mathrm{e}^{-2t} \tilde{G}^2) \, \mathrm{d}\xi \ ,$$

where $\tilde{F}(\xi, t) = F(\xi \mathrm{e}^{-t/2}, t)$, $\tilde{G}(\xi, t) = G(\xi \mathrm{e}^{-t/2}, t)$. Since $(\tilde{F}, \tilde{G}) \in \mathcal{C}^1([t_0, t_1], \mathrm{X}_0^1 \times \mathrm{X}_0^0)$ (see the remark at the end of the previous section), it follows that $E_1, \mathcal{E}_1 \in \mathcal{C}^1([t_0, t_1])$. Using (2.40), we thus find

$$\begin{aligned} \dot{E}_1 &= \frac{E_1}{2} + \int_{\mathbf{R}} \frac{\mathrm{e}^t}{p(x\mathrm{e}^{t/2})} \big(FF_t - \tfrac{x}{2}FF_x + \eta \mathrm{e}^{-t}\big((FG)_t - \tfrac{x}{2}(FG)_x - FG\big) \big) \, \mathrm{d}x \ , \\ \dot{\mathcal{E}}_1 &= \frac{3\mathcal{E}_1}{2} + \int_{\mathbf{R}} \frac{\mathrm{e}^t}{p(x\mathrm{e}^{t/2})} \big(F_xF_{xt} - \tfrac{x}{2}F_xF_{xx} - \tfrac{1}{2}F_x^2 + \eta \mathrm{e}^{-t}\big(GG_t - \tfrac{x}{2}GG_x - G^2\big) \big) \, \mathrm{d}x \ . \end{aligned}$$

Applying the identities (3.2), (3.3) and the relation $F_t = G + \frac{x}{2}F_x$, we obtain the desired result after some integrations by parts. This concludes the proof of Lemma 3.2. □

We now define positive constants A_0, B_0 by

$$A_0 \,=\, 2\left(\inf_{\xi\geq 0} p(\xi)|\gamma'(\xi)|\right)^{-1} , \quad B_0 \,=\, \left(\sup_{\xi\in\mathbf{R}} p(\xi)\gamma(\xi)^2\right)^{-1} . \tag{3.5}$$

Due to (1.11), (1.12), (2.26), these quantities are well defined. Moreover, the inequality $|\gamma'(\xi)| \leq \frac{1}{2}\gamma(\xi)^2$ implies that $A_0 \geq 4B_0 > 0$. With these notations, we introduce the functional

$$S_1(t) \,=\, A_0E_0(t) + B_0\mathcal{E}_0(t) + 2E_1(t) + \mathcal{E}_1(t) , \quad t\in[t_0,t_1] .$$

Proposition 3.3. *Assume that $\eta\mathrm{e}^{-t_0}$ is sufficiently small, and that $(u,v) \in \mathcal{C}([t_0,t_1],\mathrm{Z}_t)$ is a solution of (1.16) satisfying the bound (2.20). Then $S_1 \in \mathcal{C}^1([t_0,t_1])$, $S_1(t) \geq 0$, and there exist positive constants K_5, K_6 such that, for all $t\in[t_0,t_1]$,*

$$\dot S_1(t) + \frac{1}{2}S_1(t) \,\leq\, -K_5\int_0^\infty (x^2+x^4)f^2\,\mathrm{d}x + K_6\mathrm{e}^{-t/4}(\|f\|_{\mathrm{L}^2_t}+\|g\|_{\mathrm{L}^2_t})M(t) , \tag{3.6}$$

where $M(t)^2 \,=\, \alpha(t)^2+\beta(t)^2+\|f(t)\|^2_{\mathrm{H}^1_t}+\|g(t)\|^2_{\mathrm{L}^2_t}$.

Proof. Assuming $\eta\mathrm{e}^{-t_0} \leq \min(1/2, B_0/A_0)$, one verifies that $A_0E_0(t)+B_0\mathcal{E}_0(t) \geq 0$ and $2E_1(t)+\mathcal{E}_1(t) \geq 0$ for $t\in[t_0,t_1]$. Next, we remark that $F_x = \mathrm{e}^{-t}p(x\mathrm{e}^{t/2})f$, hence $\|F_x\|_{\mathrm{L}^2} \leq C\|f\|_{\mathrm{L}^2_t}$ by (1.17), (2.26). Thus, using Lemmas 2.7 and 2.8, we deduce from Lemma 3.1 that

$$\dot E_0(t) + \frac{E_0(t)}{2} \,\leq\, \int_{\mathbf{R}}\left(-F_x^2 + \frac{\mathrm{e}^t}{2}\gamma'(x\mathrm{e}^{t/2})F^2\right)\mathrm{d}x + C\mathrm{e}^{-t/4}(\|f\|_{\mathrm{L}^2_t}+\|g\|_{\mathrm{L}^2_t})M(t) .$$

Similarly, using the bound $|\mathrm{e}^{t/2}\gamma(x\mathrm{e}^{t/2})F_xG| \leq \frac{1}{2}(G^2+\mathrm{e}^t\gamma(x\mathrm{e}^{t/2})^2F_x^2)$, we obtain

$$\dot{\mathcal{E}}_0(t) + \frac{\mathcal{E}_0(t)}{2} \,\leq\, \frac{1}{2}\int_{\mathbf{R}}\left(-G^2+F_x^2+\mathrm{e}^t\gamma(x\mathrm{e}^{t/2})^2F_x^2\right)\mathrm{d}x + C\mathrm{e}^{-t/4}\|g\|_{\mathrm{L}^2_t}M(t) .$$

Finally, applying Lemma 2.7 and Lemma 2.8 again, we deduce from Lemma 3.2 that

$$\begin{aligned} 2\dot E_1(t) + \dot{\mathcal{E}}_1(t) + E_1(t) + \frac{1}{2}\mathcal{E}_1(t) \,&\leq\, C\mathrm{e}^{-t/4}(\|f\|_{\mathrm{L}^2_t}+\|g\|_{\mathrm{L}^2_t})M(t) \\ &+ \int_{\mathbf{R}}\frac{\mathrm{e}^t}{p(x\mathrm{e}^{t/2})}\left(-F_x^2-G^2+F^2+2\nu\mathrm{e}^{-t/2}(2F+G)G_x\right)\mathrm{d}x . \end{aligned}$$

The last term in the right-hand side is bounded with the help of (2.17), (2.26), and Lemma 2.7:

$$\begin{aligned} \int_{\mathbf{R}}\frac{\mathrm{e}^{t/2}}{p(x\mathrm{e}^{t/2})}|(2F+G)G_x|\,\mathrm{d}x &\leq C\left(\int_{\mathbf{R}}\frac{\mathrm{e}^t(F^2+G^2)}{p(x\mathrm{e}^{t/2})}\,\mathrm{d}x\right)^{1/2}\left(\int_{\mathbf{R}}\mathrm{e}^{-2t}p(x\mathrm{e}^{t/2})g^2\,\mathrm{d}x\right)^{1/2} \\ &\leq C\mathrm{e}^{-t/2}(\|f\|_{\mathrm{L}^2_t}+\|g\|_{\mathrm{L}^2_t})\|g\|_{\mathrm{L}^2_t} . \end{aligned}$$

Combining these estimates and using (2.35), (3.5) together with the bound $\gamma'(\xi)\leq 0$ for $\xi\leq 0$, we obtain

$$\begin{aligned} \dot S_1(t) + \frac{S_1(t)}{2} \,\leq\, &-\frac{B_0}{2}\int_{\mathbf{R}}(7F_x^2+G^2)\,\mathrm{d}x - \int_{\mathbf{R}}\frac{\mathrm{e}^t}{p(x\mathrm{e}^{t/2})}\left(\frac{1}{2}F_x^2+G^2\right)\mathrm{d}x \\ &+ C\mathrm{e}^{-t/4}(\|f\|_{\mathrm{L}^2_t}+\|g\|_{\mathrm{L}^2_t})M(t) \end{aligned}$$

for all $t \in [t_0, t_1]$, and (3.6) follows using (2.17), (2.26). This concludes the proof of Proposition 3.3. □

In the rest of this section, we introduce three pairs of weighted functionals $E_i, \mathcal{E}_i$ $(i = 2, 3, 4)$ to control the solutions (f, g) of (2.14) in the space Z_t. To each pair will correspond a different weight function $p_i : \mathbf{R} \to \mathbf{R}_+$. To define the weight p_2, we choose a smooth function $\chi_2 : \mathbf{R} \to (0, 1]$ satisfying $\chi_2(\xi) = 2\kappa/\gamma_- < 1$ for $\xi \le -1$ and $\chi_2(\xi) = 1$ for $\xi \ge 0$. We set $\gamma_2 = \chi_2\gamma$. The weight $p_2 : \mathbf{R} \to \mathbf{R}_+$ is then the (unique) solution of the differential problem

$$p_2'(\xi) \,=\, \gamma_2(\xi) p_2(\xi) \;, \quad \xi \in \mathbf{R} \;, \quad \lim_{\xi \to +\infty} \frac{p_2(\xi)}{\xi^2} \,=\, 1 \;. \tag{3.7}$$

Clearly, $p_2(\xi) = p(\xi)$ for $\xi \ge 0$, and there exists $C \ge 1$ such that $C^{-1}\mathrm{e}^{2\kappa\xi} \le p_2(\xi) \le C\mathrm{e}^{2\kappa\xi}$ for $\xi \le 0$. In particular, we have for all $u \in \mathrm{L}^2_t$

$$\int_{-\infty}^0 p_2(x\mathrm{e}^{t/2}) u(x)^2 \,\mathrm{d}x + \int_0^\infty \mathrm{e}^{-t} p_2(x\mathrm{e}^{t/2}) u(x)^2 \,\mathrm{d}x \,\le\, C\|u\|^2_{\mathrm{L}^2_t} \;. \tag{3.8}$$

We now define the functionals

$$\begin{aligned} E_2(t) \,&=\, \int_{\mathbf{R}} \mathrm{e}^{-t} p_2(x\mathrm{e}^{t/2}) \left(\frac{1}{2} f^2 + \eta \mathrm{e}^{-t} fg \right) \mathrm{d}x \;, \\ \mathcal{E}_2(t) \,&=\, \frac{1}{2} \int_{\mathbf{R}} \mathrm{e}^{-t} p_2(x\mathrm{e}^{t/2}) \left(f_x^2 + \eta \mathrm{e}^{-t} g^2 \right) \mathrm{d}x \end{aligned} \tag{3.9}$$

together with $S_2(t) = 2E_2(t) + \mathcal{E}_2(t)$.

Proposition 3.4. *Assume that $\eta \mathrm{e}^{-t_0} \le 1/8$ and that $(u, v) \in \mathcal{C}([t_0, t_1], \mathrm{Z}_t)$ is a solution of (1.16) satisfying the bound (2.20). Then $S_2 \in \mathcal{C}^1([t_0, t_1])$ and there exist positive constants K_7, K_8 such that, for all $t \in [t_0, t_1]$,*

$$S_2(t) \,\ge\, \frac{1}{4} \int_{\mathbf{R}} \mathrm{e}^{-t} p_2(x\mathrm{e}^{t/2}) (f^2 + f_x^2 + \eta \mathrm{e}^{-t} g^2) \,\mathrm{d}x \;, \tag{3.10}$$

and

$$\begin{aligned} \dot{S}_2 + \frac{1}{2} S_2 \,\le\, &- K_7 \left(\int_0^\infty x^2 \big(f_x^2 + g^2 \big) \,\mathrm{d}x + \int_{-\infty}^0 \mathrm{e}^{2\kappa x \mathrm{e}^{t/2}} f^2 \,\mathrm{d}x \right) \\ &+ K_8 \left(\int_0^\infty x^2 f^2 \,\mathrm{d}x + \mathrm{e}^{-t/4} (\|f\|_{\mathrm{L}^2_t} + \|g\|_{\mathrm{L}^2_t}) M(t) \right) \;. \end{aligned} \tag{3.11}$$

Proof. Since $2\eta \mathrm{e}^{-t}|fg| \le 4\eta \mathrm{e}^{-t} f^2 + \frac{1}{4}\eta \mathrm{e}^{-t} g^2$ and $\eta \mathrm{e}^{-t} \le 1/8$, the lower bound (3.10) is obvious. To compute the time derivative of E_2, we note that

$$E_2(t) \,=\, \int_{\mathbf{R}} \mathrm{e}^{-3t/2} p_2(\xi) \left(\frac{1}{2} \tilde{f}^2 + \eta \mathrm{e}^{-t} \tilde{f} \tilde{g} \right) \mathrm{d}\xi \;,$$

where $\tilde{f}(\xi, t) = f(\xi \mathrm{e}^{-t/2}, t)$, $\tilde{g}(\xi, t) = g(\xi \mathrm{e}^{-t/2}, t)$. If we assume that $(u(t_0), v(t_0)) \in \mathrm{H}^2_{t_0} \times \mathrm{H}^1_{t_0}$, then (as in Lemma 2.1) $(\tilde{f}, \tilde{g}) \in \mathcal{C}([t_0, t_1], \mathrm{H}^2_0 \times \mathrm{H}^1_0) \cap \mathcal{C}^1([t_0, t_1], \mathrm{H}^1_0 \times \mathrm{L}^2_0)$ and $\tilde{f}_t(\xi, t) = (f_t - \frac{x}{2} f_x)(\xi \mathrm{e}^{-t/2}, t)$, $\tilde{g}_t(\xi, t) = (g_t - \frac{x}{2} g_x)(\xi \mathrm{e}^{-t/2}, t)$. A direct calculation then yields

$$\dot{E}_2(t) \,=\, \int_{\mathbf{R}} \mathrm{e}^{-t} p_2(x\mathrm{e}^{t/2}) \left(f f_t - \frac{x}{2} f f_x - \frac{3}{4} f^2 + \eta \mathrm{e}^{-t} \left((fg)_t - \frac{x}{2} (fg)_x - \frac{5}{2} fg \right) \right) \mathrm{d}x \;.$$

Applying the identity

$$\begin{aligned} &ff_t + \eta \mathrm{e}^{-t}\big((fg)_t - \tfrac{x}{2}(fg)_x - 4fg - g^2\big) \\ &\quad = ff_{xx} + (\tfrac{x}{2} + \mathrm{e}^{t/2}\gamma(x\mathrm{e}^{t/2}))ff_x + \tfrac{3}{2}f^2 + \nu\gamma(x\mathrm{e}^{t/2})fg + 2\nu\mathrm{e}^{-t/2}fg_x + fr \ , \end{aligned} \tag{3.12}$$

which follows from (2.14), and integrating by parts, we obtain

$$\begin{aligned} \dot{E}_2 \ = \ \frac{3E_2}{2} + \int_{\mathbf{R}} \mathrm{e}^{-t} p_2(x\mathrm{e}^{t/2}) \bigg(&-f_x^2 + \frac{1}{2}\mathrm{e}^t \Gamma_2(x\mathrm{e}^{t/2}) f^2 + \eta \mathrm{e}^{-t} g^2 \\ &+ \nu(\gamma - 2\gamma_2)(x\mathrm{e}^{t/2})fg - 2\nu\mathrm{e}^{-t/2} f_x g + fr \bigg) \, \mathrm{d}x \ , \end{aligned} \tag{3.13}$$

where $\Gamma_2 = \gamma_2' - \gamma' - \gamma_2(\gamma - \gamma_2)$. As is easily verified, the right-hand side of (3.13) is a continuous function of the initial data $(u(t_0), v(t_0))$ in the topology of Z_{t_0}, uniformly in $t \in [t_0, t_1]$. Therefore, using a density argument as in the proof of Lemma 2.3, we conclude that $E_2 \in \mathcal{C}^1([t_0, t_1])$ and that (3.13) holds in the general case where $(u(t_0), v(t_0)) \in \mathrm{Z}_{t_0}$ only.

In a similar way, we obtain for regular data

$$\dot{\mathcal{E}}_2 \ = \ \int_{\mathbf{R}} \mathrm{e}^{-t} p_2(x\mathrm{e}^{t/2}) \left(f_x f_{xt} - \frac{x}{2} f_x f_{xx} - \frac{3}{4} f_x^2 + \eta \mathrm{e}^{-t} \left(g g_t - \frac{x}{2} g g_x - \frac{5}{4} g^2 \right) \right) \mathrm{d}x \ .$$

Using the relation $f_t = g + \frac{x}{2} f_x + \frac{3}{2} f$ as well as the identity

$$\begin{aligned} -f_{xx} f_t + \eta \mathrm{e}^{-t}\big(g g_t - \tfrac{x}{2} g g_x - \tfrac{5}{2} g^2 \big) \ &= \ -(1 - \nu\gamma(x\mathrm{e}^{t/2}))g^2 + 2\nu\mathrm{e}^{-t/2} g g_x \\ &\quad - \tfrac{x}{2} f_x f_{xx} - \tfrac{3}{2} f f_{xx} + \mathrm{e}^{t/2}\gamma(x\mathrm{e}^{t/2}) f_x g + gr \ , \end{aligned} \tag{3.14}$$

which follows from (2.14), we obtain after integrating by parts

$$\dot{\mathcal{E}}_2 \ = \ \frac{5\mathcal{E}_2}{2} + \int_{\mathbf{R}} \mathrm{e}^{-t} p_2(x\mathrm{e}^{t/2}) \big(-g^2 + (\nu g^2 + \mathrm{e}^{t/2} f_x g)(\gamma - \gamma_2)(x\mathrm{e}^{t/2}) + gr \big) \, \mathrm{d}x \ . \tag{3.15}$$

By the same density argument, $\mathcal{E}_2 \in \mathcal{C}^1([t_0, t_1])$ and (3.15) holds for all solutions (u, v) of (1.16) in Z_t.

We now estimate the right-hand side of (3.13). Since $|(\gamma - 2\gamma_2)(\xi)| \le \gamma(\xi)$ for $\xi \in \mathbf{R}$ and $\mathrm{e}^{-t} p_2(x\mathrm{e}^{t/2})\gamma(x\mathrm{e}^{t/2}) \le C\mathrm{e}^{-t/2}(1+x)$ for $x \ge 0$, we obtain with the help of (3.8)

$$\int_{\mathbf{R}} \mathrm{e}^{-t} p_2(x\mathrm{e}^{t/2}) |(\gamma - 2\gamma_2)(x\mathrm{e}^{t/2}) fg| \, \mathrm{d}x \ \le \ C\mathrm{e}^{-t/2} \|f\|_{\mathrm{L}^2_t} \|g\|_{\mathrm{L}^2_t} \ . \tag{3.16}$$

Remarking that $\Gamma_2(\xi) = 0$ for $\xi \ge 0$, we deduce from (3.8), (3.13), (3.16), and Lemma 2.6 that

$$\begin{aligned} \dot{E}_2 + \frac{1}{2} E_2 \le \int_0^\infty \mathrm{e}^{-t} p_2(x\mathrm{e}^{t/2})(f^2 - f_x^2) \, \mathrm{d}x + \frac{1}{2} \int_{-\infty}^0 p_2(x\mathrm{e}^{t/2}) \Gamma_2(x\mathrm{e}^{t/2}) f^2 \, \mathrm{d}x \\ + C\mathrm{e}^{-t/4}(\|f\|_{\mathrm{L}^2_t} + \|g\|_{\mathrm{L}^2_t}) M(t) \ . \end{aligned} \tag{3.17}$$

Since $\Gamma_2(\xi) \to -2\kappa(\gamma_- - 2\kappa) = -2\kappa c_*$ as $\xi \to -\infty$, we can write $\Gamma_2(\xi) \le -\kappa c_* + \Gamma_2^*(\xi)$ for all $\xi \le 0$, where Γ_2^* is a bounded nonnegative function with support in a compact

interval $[-A,0]$. Applying Lemma 2.4, we thus obtain

$$\begin{aligned}(3.18)\qquad &\int_{-\infty}^0 p_2(x\mathrm{e}^{t/2})\Gamma_2(x\mathrm{e}^{t/2})f^2\,\mathrm{d}x+\kappa c_*\int_{-\infty}^0 p_2(x\mathrm{e}^{t/2})f^2\,\mathrm{d}x\\ &\qquad\le \mathrm{e}^{-t/2}\sup_{x\ge -A\mathrm{e}^{-t/2}} f(x,t)^2\int_{-A}^0 p_2(\xi)\Gamma_2^*(\xi)\,\mathrm{d}\xi \,\le\, C\mathrm{e}^{-t/2}\|f\|_{\mathrm{H}^1_t}^2\ .\end{aligned}$$

Similarly, remarking that $\gamma_2(\xi)=\gamma(\xi)$ for $\xi\ge 0$, we deduce from (3.8), (3.15), and Lemma 2.6 that

$$(3.19)\quad \dot{\mathcal{E}}_2+\frac{1}{2}\mathcal{E}_2\,\le\,\int_0^\infty \mathrm{e}^{-t}p_2(x\mathrm{e}^{t/2})(\tfrac{3}{2}f_x^2-g^2)\,\mathrm{d}x+C\mathrm{e}^{-t/4}(\|f\|_{\mathrm{L}^2_t}+\|g\|_{\mathrm{L}^2_t})M(t)\ .$$

Combining (3.17), (3.18), (3.19), and using (2.26), (3.7), we obtain (3.11). This concludes the proof of Proposition 3.4. □

The construction of our next functionals E_3, $\mathcal{E}_3$ is one of the main difficulties in the proof of Theorem 1.2. The aim is to control the quantity

$$\int_{-\infty}^0 \mathrm{e}^{2\kappa x\mathrm{e}^{t/2}}\left(f^2+f_x^2+\eta\mathrm{e}^{-t}g^2\right)\,\mathrm{d}x+\int_0^\infty\left(f^2+f_x^2+\eta\mathrm{e}^{-t}g^2\right)\,\mathrm{d}x\ ,$$

which is part of the norm of (f,g) in Z_t. A natural idea is to define E_3, $\mathcal{E}_3$ by the formulas (3.9) with $\mathrm{e}^{-t}p_2(x\mathrm{e}^{t/2})$ replaced by $p_3(x\mathrm{e}^{t/2})$, where $p_3(\xi)=\mathcal{O}(\mathrm{e}^{2\kappa\xi})$ as $\xi\to-\infty$ and $p_3(\xi)\to 1$ as $\xi\to+\infty$. However, we are not able to estimate properly the time derivative of these functionals without including in $\mathcal{E}_3$ an additional term of the form

$$\int_{\mathbf{R}} p_3(x\mathrm{e}^{t/2})\lambda(x\mathrm{e}^{t/2})\gamma(x\mathrm{e}^{t/2})(\nu f_x^2-\eta\mathrm{e}^{-t/2}f_xg)\,\mathrm{d}x\ ;$$

see (3.25) below. With this modification, the derivative of $\mathcal{E}_3$ contains a quadratic form $Q(x,t)$ depending on the functions λ and p_3; see (3.30). As we shall show, the evolution of $E_3,\mathcal{E}_3$ can then be controlled provided $Q(x,t)$ is positive definite.

We now construct positive functions λ, p_3 so that the quadratic form $Q(x,t)$ in (3.30) is positive definite. First of all, since $\gamma_-=c_*+2\kappa>c_*$ and $\nu c_*<1$ by (1.9), we can define

$$(3.20)\qquad \lambda_-\,=\,\left(\frac{\gamma_-^2}{c_*^2}-\nu\gamma_-\right)^{-1}\,>\,0\ .$$

For later use, we remark that

$$(3.21)\qquad \lambda_-(1-\nu c_*)\,<\,(c_*/\gamma_-)^2<1\ ,\quad\text{and}\quad \lambda_-\gamma_-\,<\,\nu/\eta\ .$$

Next, in view of (1.11), (1.12), we can choose $\xi_3>0$ sufficiently large so that

$$(3.22)\qquad \gamma(\xi_3)\,<\,c_*\lambda_-\ ,\quad \nu\gamma(\xi_3)\,\le\,\frac{1}{2}\ ,\quad \gamma'(\xi)\,\le\,-\frac{1}{4}\gamma(\xi)^2\quad\text{for all}\quad \xi\ge\xi_3\ .$$

Remark that the first inequality in (3.22) is automatically satisfied if $\lambda_-\ge 1$, since $\gamma(0)=c_*$ and γ is nonincreasing. The conditions (3.21), (3.22) imply that there exists a smooth function $\lambda:\mathbf{R}\to\mathbf{R}_+$ satisfying $\lambda(\xi)=\lambda_-$ if $\xi\le 0$, $\lambda(\xi)=1$ if $\xi\ge\xi_3$, $(\lambda\gamma)'(\xi)\le 0$ for all $\xi\in\mathbf{R}$, and

$$(3.23)\qquad \lambda(\xi)\big((1-\nu\gamma(\xi))^2+\eta\gamma(\xi)^2\big)\,\le\,1\ ,\quad \xi\in[0,\xi_3]\ .$$

Indeed, assume first that $\lambda_- < 1$. Then the first inequality in (3.22) ensures that λ can be constructed so that $\lambda'(\xi) \geq 0$ and $(\lambda\gamma)'(\xi) \leq 0$ for $\xi \in [0, \xi_3]$. In particular, $\lambda(\xi) \leq 1$ for all $\xi \in \mathbf{R}$. On the other hand, we observe that the function

$$\Omega(\gamma) \,=\, (1-\nu\gamma)^2 + \eta\gamma^2 \,\equiv\, 1 - 2\nu\gamma + \frac{\nu\gamma^2}{c_*}$$

is nonincreasing for $\gamma \leq c_*$, with $\Omega(0) = 1$ and $\Omega(c_*) = 1 - \nu c_* > 0$. Since $\gamma(\xi) \leq c_*$ for $\xi \geq 0$, we have $\lambda(\xi)\Omega(\gamma(\xi)) \leq \Omega(0) = 1$ for $\xi \in [0, \xi_3]$, which is (3.23). Assume now that $\lambda_- \geq 1$, and choose λ so that $\lambda'(\xi) \leq 0$ for all $\xi \in \mathbf{R}$. Then the condition $(\lambda\gamma)'(\xi) \leq 0$ is automatically satisfied. Morevover, since $\lambda_-\Omega(\gamma(0)) < (c_*/\gamma_-)^2 < 1$ by (3.21), it is sufficient to assume that $\lambda(\xi)$ decays rapidly enough from λ_- to 1 (as ξ varies from 0 to ξ_3) so that (3.23) is satisfied.

We next define the weight function p_3. Let $\chi_3 : \mathbf{R} \to (0, 1]$ be a smooth function satisfying $\chi_3(\xi) = 2\kappa/\gamma_- < 1$ for $\xi \leq -1$, $\chi_3(\xi) = 1$ for $\xi \in [0, \xi_3]$, and $\chi_3(\xi) = 0$ for $\xi \geq \xi_3 + 1$. We also assume that $\xi\chi_3'(\xi) \leq 0$ for all $\xi \in \mathbf{R}$. We set $\gamma_3 = \chi_3\gamma$, and define the weight function $p_3 : \mathbf{R} \to \mathbf{R}_+$ as the (unique) solution of the differential problem

$$p_3'(\xi) \,=\, \gamma_3(\xi)p_3(\xi)\,, \quad \xi \in \mathbf{R}\,, \quad \lim_{\xi\to+\infty} p_3(\xi) \,=\, 1\,. \tag{3.24}$$

Clearly, there exists $C \geq 1$ such that $C^{-1} \leq p_3(\xi) \leq C$ for $\xi \geq 0$ and $C^{-1}\mathrm{e}^{2\kappa\xi} \leq p_3(\xi) \leq C\mathrm{e}^{2\kappa\xi}$ for $\xi \leq 0$.

With these definitions, we now introduce the functionals

$$\begin{aligned} E_3(t) \,&=\, \int_{\mathbf{R}} p_3(x\mathrm{e}^{t/2}) \left(\frac{1}{2}f^2 + \eta\mathrm{e}^{-t}fg\right) \mathrm{d}x\,, \\ \mathcal{E}_3(t) \,&=\, \frac{1}{2}\int_{\mathbf{R}} p_3(x\mathrm{e}^{t/2})\big(f_x^2 + \eta\mathrm{e}^{-t}g^2 + 2(\lambda\gamma)(x\mathrm{e}^{t/2})(\nu f_x^2 - \eta\mathrm{e}^{-t/2}f_xg)\big)\,\mathrm{d}x\,, \end{aligned} \tag{3.25}$$

together with $S_3(t) = KE_3(t) + \mathcal{E}_3(t)$, where $K = 3 + 4\nu\|\lambda\gamma\|_{\mathrm{L}^\infty}$.

Proposition 3.5. *Assume that $\eta\mathrm{e}^{-t_0}$ is sufficiently small, and that $(u, v) \in \mathcal{C}([t_0, t_1], \mathrm{Z}_t)$ is a solution of (1.16) satisfying the bound (2.20). Then $S_3 \in \mathcal{C}^1([t_0, t_1])$, and there exist positive constants K_9, K_{10} such that, for all $t \in [t_0, t_1]$,*

$$S_3(t) \,\geq\, \frac{1}{8}\int_{\mathbf{R}} p_3(x\mathrm{e}^{t/2})(f^2 + f_x^2 + \eta\mathrm{e}^{-t}g^2)\,\mathrm{d}x\,, \tag{3.26}$$

and

$$\begin{aligned} \dot{S}_3(t) + \frac{1}{2}S_3(t) \,\leq\, &- K_9\int_{\mathbf{R}} p_3(x\mathrm{e}^{t/2})(g^2 + f_x^2 + \mathrm{e}^t\gamma(x\mathrm{e}^{t/2})^2 f_x^2)\,\mathrm{d}x \\ &+ K_{10}\left(\int_{-\infty}^0 \mathrm{e}^{2\kappa x\mathrm{e}^{t/2}} f^2\,\mathrm{d}x + \int_0^\infty x^2 f_x^2\,\mathrm{d}x\right) \\ &+ K_{10}\big(\mathrm{e}^{-t/4}(\|f\|_{\mathrm{H}^1_t} + \|g\|_{\mathrm{L}^2_t})M(t) + \mathrm{e}^{-t/2}M(t)^2\big)\,. \end{aligned} \tag{3.27}$$

Proof. Since $|Kfg| \leq \frac{1}{8}g^2 + 2K^2f^2$ and $|\mathrm{e}^{-t/2}\lambda\gamma f_xg| \leq \frac{1}{4}\mathrm{e}^{-t}g^2 + (\lambda\gamma)^2 f_x^2$, we have

$$S_3 \geq \int_{\mathbf{R}} p_3(x\mathrm{e}^{t/2}) \left(\left(\frac{K}{2} - 2K^2\eta \mathrm{e}^{-t} \right) f^2 + \frac{1}{8}\eta \mathrm{e}^{-t} g^2 + \frac{1}{2} f_x^2 \big(1 + 2\nu\lambda\gamma - 2\eta\lambda^2\gamma^2\big)(x\mathrm{e}^{t/2}) \right) \mathrm{d}x \ .$$

Assuming that $\eta \mathrm{e}^{-t_0} \leq (8K)^{-1}$ and noting that $\nu - \eta\lambda\gamma \geq \nu - \eta\lambda_-\gamma_- > 0$ by (3.21), we obtain (3.26).

Next, proceeding as in the proof of Proposition 3.4, we show that $E_3 \in \mathcal{C}^1([t_0, t_1])$ and that

$$\dot{E}_3 = \frac{5E_3}{2} + \int_{\mathbf{R}} p_3(x\mathrm{e}^{t/2}) \big(-f_x^2 + \mathrm{e}^{t/2}(\gamma - \gamma_3)(x\mathrm{e}^{t/2}) f f_x + \eta \mathrm{e}^{-t} g^2 + \nu(\gamma - 2\gamma_3)(x\mathrm{e}^{t/2}) fg - 2\nu \mathrm{e}^{-t/2} f_x g + fr \big) \mathrm{d}x \tag{3.28}$$

for $t \in [t_0, t_1]$. The analysis of $\mathcal{E}_3$ is more complicated due to the additional term $2\lambda\gamma(\nu f_x^2 - \eta \mathrm{e}^{-t/2} f_x g)$. First, assuming that the initial data are regular, we obtain by a direct calculation

$$\dot{\mathcal{E}}_3 = \int_{\mathbf{R}} p_3(x\mathrm{e}^{t/2}) \big((1 + 2\nu(\lambda\gamma)(x\mathrm{e}^{t/2}))(f_x f_{xt} - \tfrac{x}{2} f_x f_{xx} - \tfrac{1}{4} f_x^2) + \eta \mathrm{e}^{-t}(g g_t - \tfrac{x}{2} g g_x - \tfrac{3}{4} g^2) - \eta \mathrm{e}^{-t/2}(\lambda\gamma)(x\mathrm{e}^{t/2})((f_x g)_t - \tfrac{x}{2}(f_x g)_x - f_x g) \big) \mathrm{d}x \ .$$

Using the relation $f_t = g + \frac{x}{2} f_x + \frac{3}{2} f$ together with the identities (3.14) and

$$\begin{aligned} 2\nu f_x f_{xt} - \eta \mathrm{e}^{-t/2} \big((f_x g)_t - \tfrac{x}{2}(f_x g)_x - \tfrac{9}{2} f_x g - g g_x \big) &= \mathrm{e}^{t/2}(1 - \nu\gamma(x\mathrm{e}^{t/2})) f_x g \\ &\quad - \mathrm{e}^{t/2} f_x f_{xx} - \mathrm{e}^{t} \gamma(x\mathrm{e}^{t/2}) f_x^2 + \nu x f_x f_{xx} + 4\nu f_x^2 - \mathrm{e}^{t/2} f_x r \ , \end{aligned} \tag{3.29}$$

which follow from (2.14), we obtain after integrating by parts

$$\dot{\mathcal{E}}_3 = \frac{7\mathcal{E}_3}{2} + \int_{\mathbf{R}} p_3(x\mathrm{e}^{t/2}) \big((g - \mathrm{e}^{t/2}(\lambda\gamma)(x\mathrm{e}^{t/2}) f_x) r - Q(x,t)[\mathrm{e}^{t/2}\gamma(x\mathrm{e}^{t/2}) f_x, g] \big) \mathrm{d}x \ , \tag{3.30}$$

where $Q(x,t)$ is the quadratic form defined by

$$\begin{aligned} Q(x,t)[z_1, z_2] = {} & z_1^2 \big(\lambda - \tfrac{1}{2}\lambda\gamma^{-1}(\gamma_3 + \mu) \big)(x\mathrm{e}^{t/2}) - z_1 z_2 \big(1 - \chi_3 + \lambda(1 - \nu\gamma) \big)(x\mathrm{e}^{t/2}) \\ & + z_2^2 \big(1 + \nu(\gamma_3 - \gamma) - \tfrac{1}{2}\eta\lambda\gamma(\gamma_3 + \mu) \big)(x\mathrm{e}^{t/2}) \ , \quad (z_1, z_2) \in \mathbf{R}^2 \ , \end{aligned}$$

and $\mu = (\lambda\gamma)'/(\lambda\gamma) \leq 0$. By density, (3.30) holds for all solutions (u, v) of (1.16) in Z_t.

Applying Lemma 2.6 and recalling that $K = 3 + 4\nu\|\lambda\gamma\|_{\mathrm{L}^\infty}$, we deduce from (3.28), (3.30) that

$$\begin{aligned} \dot{S}_3(t) + \frac{1}{2} S_3(t) \leq {} & \int_{\mathbf{R}} p_3(x\mathrm{e}^{t/2}) \big(-f_x^2 + \tfrac{3}{2} K f^2 - Q(x,t)[\mathrm{e}^{t/2}\gamma(x\mathrm{e}^{t/2}) f_x, g] \\ & - \mathrm{e}^{t/2}(\lambda\gamma)(x\mathrm{e}^{t/2}) f_x r + \nu K(\gamma - 2\gamma_3)(x\mathrm{e}^{t/2}) fg + K\mathrm{e}^{t/2}(\gamma - \gamma_3)(x\mathrm{e}^{t/2}) f f_x \big) \mathrm{d}x \\ & + C\mathrm{e}^{-t/4}(\|f\|_{\mathrm{H}^1_t} + \|g\|_{\mathrm{L}^2_t}) M(t) \ . \end{aligned} \tag{3.31}$$

We shall prove below that there exists $Q_0 > 0$ such that, for all $(x,t) \in \mathbf{R} \times \mathbf{R}_+$,

$$Q(x,t)[z_1, z_2] \geq Q_0(z_1^2 + z_2^2) \ , \quad (z_1, z_2) \in \mathbf{R}^2 \ . \tag{3.32}$$

Assuming for a while that (3.32) holds, and using Lemma 2.6 together with the inequalities

$$K\mathrm{e}^{t/2}\gamma(1-\chi_3)ff_x - \mathrm{e}^{t/2}\lambda\gamma f_x r \,\le\, \frac{Q_0}{2}\mathrm{e}^t\gamma^2 f_x^2 + \frac{K^2}{Q_0}f^2 + \frac{1}{Q_0}\lambda^2 r^2 \,,$$
$$\nu K\gamma(1-2\chi_3)fg \,\le\, \frac{Q_0}{2}g^2 + \frac{\nu^2K^2}{2Q_0}\gamma^2 f^2 \,,$$

we deduce from (3.31) that

(3.33)
$$\dot S_3(t) + \frac12 S_3(t) \,\le\, \int_{\mathbf{R}} p_3(x\mathrm{e}^{t/2})\left(-f_x^2 - \frac{Q_0}{2}(g^2 + \mathrm{e}^t\gamma(x\mathrm{e}^{t/2})^2 f_x^2)\right)\mathrm{d}x$$
$$+ C\left(\int_{\mathbf{R}} p_3(x\mathrm{e}^{t/2})f^2\,\mathrm{d}x + \mathrm{e}^{-t/2}M(t)^2 + \mathrm{e}^{-t/4}(\|f\|_{\mathrm{H}^1_t} + \|g\|_{\mathrm{L}^2_t})M(t)\right) \,.$$

The estimate (3.27) is then a straightforward consequence of (3.33) and of the Hardy-type inequality

$$\int_{\mathbf{R}} p_3(x\mathrm{e}^{t/2})f^2\,\mathrm{d}x \,\le\, C\left(\int_{-\infty}^0 \mathrm{e}^{2\kappa x\mathrm{e}^{t/2}}f^2\,\mathrm{d}x + \int_0^\infty x^2 f_x^2\,\mathrm{d}x\right) \,.$$

It remains to prove the property (3.32), namely,

$$\big(1-\chi_3+\lambda(1-\nu\gamma)\big)^2 \,<\, 4\big(\lambda - \tfrac12\lambda\gamma^{-1}(\gamma_3+\mu)\big)\big(1+\nu(\gamma_3-\gamma) - \tfrac12\eta\lambda\gamma(\gamma_3+\mu)\big)$$

for all $\xi\in[-\infty,+\infty]$. Expanding the products in both sides, we rewrite this condition in the equivalent form

$$(1-\chi_3)^2(1+2\nu\lambda\gamma-\eta\lambda^2\gamma^2) - 2\lambda + \lambda^2\big((1-\nu\gamma)^2+\eta\gamma^2\big)$$
$$< \eta\lambda^2\mu^2 - 2\lambda^2\eta\mu(\gamma-\gamma_3) - 2\lambda\gamma^{-1}\mu(1-\nu(\gamma-\gamma_3)) \,. \tag{3.34}$$

To prove (3.34), we first remark that the right-hand side is positive, since $\mu\le 0$, $\gamma-\gamma_3\ge 0$, and $1-\nu(\gamma-\gamma_3)\ge 1-\nu c_*>0$. We also recall that $1+2\nu\lambda\gamma-\eta\lambda^2\gamma^2\ge 1$, since $\nu-\eta\lambda\gamma>0$ (the last inequality follows from (3.21) and the fact that $(\lambda\gamma)'\le 0$). We now distinguish three cases according to whether $\xi\le 0$, $\xi\in[0,\xi_3]$, or $\xi\ge\xi_3$.

(1) If $\xi\le 0$, then $\lambda=\lambda_-$ and $1-\chi_3\le c_*/\gamma_-$; hence it is sufficient to verify the stronger condition

$$\frac{c_*^2}{\gamma_-^2}(1+2\nu\lambda_-\gamma-\eta\lambda_-^2\gamma^2) - 2\lambda_- + \lambda_-^2\big((1-\nu\gamma)^2+\eta\gamma^2\big) \,<\, 0 \tag{3.35}$$

for all $\gamma\in[c_*,\gamma_-]$. Let $\Psi(\gamma)$ denote the left-hand side of (3.35), considered as a function of γ. Using (3.21) and the relation $\nu^2+\eta=\nu/c_*$, it is not difficult to verify that Ψ is convex and that

$$\Psi(\gamma_-) \,=\, -\lambda_-^2(1-\nu c_*)\left(\frac{\gamma_-^2}{c_*^2}-1\right) \,<\, 0 \,, \quad \Psi'(c_*) \,=\, \frac{2c_*^2\lambda_-}{\gamma_-^2}(\nu-\eta c_*\lambda_-) \,>\, 0 \,.$$

Since $\Psi''>0$, it follows that $\Psi'(\gamma)\ge\Psi'(c_*)>0$ for all $\gamma\ge c_*$, hence $\Psi(\gamma)\le\Psi(\gamma_-)<0$ for all $\gamma\in[c_*,\gamma_-]$, which is the desired inequality.

(2) If $\xi\in[0,\xi_3]$, then $\chi_3=1$; hence the left-hand side of (3.34) is negative by (3.23).

(3) If $\xi \geq \xi_3$, then $\lambda = 1$, $1-\chi_3 \leq 1$, hence the left-hand side of (3.34) is bounded from above by $\nu^2\gamma^2$. Neglecting the first two terms in the right-hand side (which are positive) and noting that $\mu = \gamma'/\gamma \leq 0$, we arrive at the stronger condition

$$\nu^2\gamma(\xi)^2 \;\leq\; -2\frac{\gamma'(\xi)}{\gamma(\xi)^2}\big(1-\nu\gamma(\xi)\big)\;, \quad \xi \geq \xi_3\;,$$

which is satisfied by assumption on ξ_3; see (3.22). This concludes the proof of Proposition 3.5. □

Finally, we introduce our last pair of functionals

$$\begin{aligned} E_4(t) \;&=\; \int_{\mathbf{R}} \mathrm{e}^{-3t} p_4(x\mathrm{e}^{t/2}) \left(\frac{1}{2}f^2 + \eta \mathrm{e}^{-t} fg\right) \mathrm{d}x\;, \\ \mathcal{E}_4(t) \;&=\; \frac{1}{2}\int_{\mathbf{R}} \mathrm{e}^{-3t} p_4(x\mathrm{e}^{t/2})(f_x^2 + \eta \mathrm{e}^{-t} g^2)\,\mathrm{d}x\;, \end{aligned} \tag{3.36}$$

where $p_4(\xi) = p(\xi)^3$. We set $S_4 = 2E_4 + \mathcal{E}_4$.

Proposition 3.6. *Assume that $\eta \mathrm{e}^{-t_0} \leq 1/8$ and that $(u,v) \in \mathcal{C}([t_0,t_1],\mathrm{Z}_t)$ is a solution of (1.16) satisfying the bound (2.20). Then $S_4 \in \mathcal{C}^1([t_0,t_1])$ and there exist positive constants K_{11}, K_{12}, such that for $t \in [t_0,t_1]$,*

$$S_4(t) \;\geq\; \frac{1}{4}\int_{\mathbf{R}} \mathrm{e}^{-3t} p_4(x\mathrm{e}^{t/2})(f^2 + f_x^2 + \eta \mathrm{e}^{-t} g^2)\,\mathrm{d}x \tag{3.37}$$

and

$$\begin{aligned} \dot S_4(t) + \frac{1}{2}S_4(t) \;&\leq\; -K_{11}\int_0^\infty x^6(f_x^2+g^2)\,\mathrm{d}x \\ &\quad + K_{12}\left(\int_0^\infty (x^4f^2 + x^2 f_x^2)\,\mathrm{d}x + \mathrm{e}^{-t/4}(\|f\|_{\mathrm{H}^1_t} + \|g\|_{\mathrm{L}^2_t})M(t)\right)\;. \end{aligned} \tag{3.38}$$

Proof. The lower bound (3.37) is proved as in (3.10). Arguing as the preceding propositions, we show that $E_4, \mathcal{E}_4 \in \mathcal{C}^1([t_0,t_1])$ and that

$$\begin{aligned} \dot E_4 \;=\; -\frac{E_4}{2} + \int_{\mathbf{R}} \mathrm{e}^{-3t} p_4(x\mathrm{e}^{t/2})\big(&- f_x^2 - 2\mathrm{e}^{t/2}\gamma(x\mathrm{e}^{t/2}) f f_x + \eta \mathrm{e}^{-t} g^2 \\ &- 5\nu\gamma(x\mathrm{e}^{t/2}) fg - 2\nu \mathrm{e}^{-t/2} f_x g + fr\big)\,\mathrm{d}x\;, \end{aligned}$$

$$\dot{\mathcal{E}}_4 \;= \frac{\mathcal{E}_4}{2} + \int_{\mathbf{R}} \mathrm{e}^{-3t} p_4(x\mathrm{e}^{t/2})\big(-g^2 - 2\gamma(x\mathrm{e}^{t/2})(\nu g^2 + \mathrm{e}^{t/2} f_x g) + gr\big)\,\mathrm{d}x\;.$$

Proceeding as in (3.16) and applying Lemma 2.6, we deduce that, for $t \in [t_0,t_1]$,

$$\begin{aligned} \dot S_4(t) + \frac{1}{2}S_4(t) \;&\leq\; C\mathrm{e}^{-t/4}(\|f\|_{\mathrm{H}^1_t} + \|g\|_{\mathrm{L}^2_t})M(t) \\ &\quad + \int_0^\infty \mathrm{e}^{-3t} p_4(x\mathrm{e}^{t/2})\big(-\tfrac{3}{2}f_x^2 - g^2 - 2\mathrm{e}^{t/2}\gamma(x\mathrm{e}^{t/2})(f_x g + 2f f_x)\big)\,\mathrm{d}x\;. \end{aligned} \tag{3.39}$$

Since $p_4(\xi) = p(\xi)^3 \leq C_0^3(1+\xi)^6$ for $\xi \geq 0$, we have

$$
\begin{aligned}
&\mathrm{e}^{-3t}(\gamma p_4)(x\mathrm{e}^{t/2})|2\mathrm{e}^{t/2}f_x g + 4\mathrm{e}^{t/2}f f_x| \\
&\qquad \le \mathrm{e}^{-3t}p_4(x\mathrm{e}^{t/2})\big(\tfrac{1}{2}g^2 + f_x^2 + 2\mathrm{e}^{2t}\gamma(x\mathrm{e}^{t/2})^4 f_x^2 + 8\mathrm{e}^{t}\gamma(x\mathrm{e}^{t/2})^2 f^2\big) \\
&\qquad \le \mathrm{e}^{-3t}p_4(x\mathrm{e}^{t/2})\big(\tfrac{1}{2}g^2 + f_x^2\big) + C\big(\mathrm{e}^{-t}(1+x\mathrm{e}^{t/2})^2 f_x^2 + \mathrm{e}^{-2t}(1+x\mathrm{e}^{t/2})^4 f^2\big)
\end{aligned}
$$

for $x \ge 0$, and the estimate (3.38) follows from (3.39). This concludes the proof of Proposition 3.6. □

We now summarize the decay properties of the four auxiliary functionals S_1, S_2, S_3, and S_4. To this end, we define

$$
S_5(t) \,=\, B_1 S_1(t) + B_2 S_2(t) + S_3(t) + S_4(t) + \frac{1}{2}\eta \mathrm{e}^{-t}\beta(t)^2 \,,
$$

where $B_2 = 1 + K_7^{-1}(K_{10} + K_{12})$, $B_1 = 1 + K_5^{-1}(K_8 B_2 + K_{12})$, and β is as in (2.10). In the proof of Theorem 1.2, we shall use the following properties of $S_5(t)$.

PROPOSITION 3.7. *There exist constants $A_1, A_3, A_4 > 0$ and $A_2 \ge 1$ such that, if $\eta \mathrm{e}^{-t_0} \le A_1$ and if $(u,v) \in \mathcal{C}([t_0,t_1],\mathrm{Z}_t)$ is a solution of (1.16) satisfying the bound (2.20), then, for all $t \in [t_0,t_1]$,*

$$
A_2^{-1} S_5(t) \,\le\, \|f(t)\|_{\mathrm{H}^1_t}^2 + \eta \mathrm{e}^{-t}\big(\beta(t)^2 + \|g(t)\|_{\mathrm{L}^2_t}^2\big) \,\le\, A_2 S_5(t) \,, \tag{3.40}
$$

and

$$
\begin{aligned}
\dot S_5(t) + \frac{1}{2}S_5(t) \,\le\, &- A_3\big(\beta(t)^2 + \|g(t)\|_{\mathrm{L}^2_t}^2 + \|f_x(t)\|_{\mathrm{L}^2_t}^2\big) \\
&+ A_4 \mathrm{e}^{-t/4}\big(\|f(t)\|_{\mathrm{H}^1_t} + \|g(t)\|_{\mathrm{L}^2_t} + \mathrm{e}^{-t/4}M(t)\big)M(t) \,,
\end{aligned} \tag{3.41}
$$

where $M(t)^2 = \alpha(t)^2 + \beta(t)^2 + \|f(t)\|_{\mathrm{H}^1_t}^2 + \|g(t)\|_{\mathrm{L}^2_t}^2$.

Proof. Since $S_1(t) \ge 0$ by Proposition 3.3, the lower bound on S_5 in (3.40) follows immediately from (3.10), (3.26), (3.37), and the properties of the weights p_2, p_3, p_4. The upper bound is proved in a similar way by using, in addition, Lemma 2.7 applied to F and G. On the other hand, we have, by Lemmas 2.3 and 2.6,

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{2}\eta \mathrm{e}^{-t}\beta(t)^2\right) + \frac{1}{4}\eta \mathrm{e}^{-t}\beta(t)^2 \,&=\, -\beta(t)^2 + \frac{3}{4}\eta \mathrm{e}^{-t}\beta(t)^2 + m(t)\beta(t) \\
&\le\, -\beta(t)^2 + C\mathrm{e}^{-t/2}M(t)^2 \,.
\end{aligned} \tag{3.42}
$$

Combining the estimates (3.6), (3.11), (3.27), (3.38), and (3.42), we thus obtain

$$
\begin{aligned}
\dot S_5(t) + \frac{1}{2}S_5(t) \,\le\, &-K_5\int_0^\infty (x^2+x^4) f^2 \,\mathrm{d}x - K_7 \int_{-\infty}^0 \mathrm{e}^{2\kappa x \mathrm{e}^{t/2}} f^2 \,\mathrm{d}x \\
&- \int_0^\infty (K_7 x^2 + K_{11} x^6)(f_x^2 + g^2)\,\mathrm{d}x - K_9 \int_{\mathbf{R}} p_3(x\mathrm{e}^{t/2})(g^2 + f_x^2)\,\mathrm{d}x - \beta(t)^2 \\
&+ C\mathrm{e}^{-t/4}\big(\|f(t)\|_{\mathrm{H}^1_t} + \|g(t)\|_{\mathrm{L}^2_t} + \mathrm{e}^{-t/4}M(t)\big)M(t) \,,
\end{aligned}
$$

from which (3.41) follows using the properties of the weight p_3. This concludes the proof of Proposition 3.7. □

A useful consequence of Proposition 3.7 is the following.

COROLLARY 3.8. *There exist constants $A_5 > 0$ and $A_6 \geq 1$ such that, if $t_0 \geq A_5$ and if $(u, v) \in \mathcal{C}([t_0, t_1], \mathrm{Z}_t)$ is a solution of (1.16) satisfying the bound (2.20), then*

$$\Phi_\eta(t, u(t), v(t)) \equiv \|u(t)\|^2_{\mathrm{H}^1_t} + \eta \mathrm{e}^{-t}\|v(t)\|^2_{\mathrm{L}^2_t} \leq A_6 \Phi_\eta(t_0, u(t_0), v(t_0)) \tag{3.43}$$

for all $t \in [t_0, t_1]$.

Proof. We introduce our last functional:

$$S_6(t) = \frac{1}{2}\alpha(t)^2 + \eta \mathrm{e}^{-t}\alpha(t)\beta(t) + S_5(t) \ , \quad t \in [t_0, t_1] \ .$$

In view of (3.40), if $\eta \mathrm{e}^{-t_0} \leq \min(A_1, A_2^{-1})$, there exists a constant $\tilde{C}_1 \geq 1$ such that, for $t \in [t_0, t_1]$,

$$\tilde{C}_1^{-1} S_6(t) \leq \alpha(t)^2 + \|f(t)\|^2_{\mathrm{H}^1_t} + \eta \mathrm{e}^{-t}\big(\beta(t)^2 + \|g(t)\|^2_{\mathrm{L}^2_t}\big) \leq \tilde{C}_1 S_6(t) \ . \tag{3.44}$$

By Lemma 2.5, it follows that

$$\tilde{C}_2^{-1} S_6(t) \leq \Phi_\eta(t, u(t), v(t)) \leq \tilde{C}_2 S_6(t) \ , \quad t \in [t_0, t_1] \ , \tag{3.45}$$

for some $\tilde{C}_2 \geq 1$. Now, since $\dot{S}_6(t) = \alpha(t)m(t) + \eta \mathrm{e}^{-t}\beta(t)^2 + \dot{S}_5(t)$ by (2.12), we deduce from (2.31) and (3.41) that

$$\dot{S}_6(t) \leq -A_3(\beta(t)^2 + \|g(t)\|^2_{\mathrm{L}^2_t}) + \tilde{C}_3 \mathrm{e}^{-t/4} M(t)^2 \ , \quad t \in [t_0, t_1] \ ,$$

for some $\tilde{C}_3 > 0$. Assuming that $\tilde{C}_3 \mathrm{e}^{-t_0/4} \leq A_3$ and using (3.44), we thus find $\dot{S}_6(t) \leq \tilde{C}_1 \tilde{C}_3 \mathrm{e}^{-t/4} S_6(t)$ for $t \in [t_0, t_1]$, hence $S_6(t) \leq \tilde{C}_4 S_6(t_0)$ for $t \in [t_0, t_1]$, where $\tilde{C}_4 = \exp(4\tilde{C}_1\tilde{C}_3)$. Combining this estimate with (3.45), we obtain (3.43). This concludes the proof of Corollary 3.8. □

4. End of the proof of Theorem 1.2. Let $t_0 \geq A_5$ and $\delta_0 \leq (2A_6)^{-1/2}$, where A_5, A_6 are as in Corollary 3.8. If $(u_0, v_0) \in \mathrm{Z}_{t_0}$ satisfies $\Phi_\eta(t_0, u_0, v_0) \leq \delta_0^2$, then the system (1.16) has a unique global solution $(u, v) \in \mathcal{C}([t_0, +\infty), \mathrm{Z}_t)$ with $(u(t_0), v(t_0)) = (u_0, v_0)$. Indeed, the local existence and uniqueness follow from Proposition 2.2, and Corollary 3.8 shows that $\Phi_\eta(t, u(t), v(t)) \leq 1/2$ as long as the solution $(u(t), v(t))$ exists. Then Proposition 2.2, with $\delta_1 = 1/\sqrt{2}$, implies that the solution $(u(t), v(t))$ is globally defined.

It remains to prove the decay estimate (1.23). Since

$$\begin{aligned}
A_4 \mathrm{e}^{-t/4}\|g\|_{\mathrm{L}^2_t} M &\leq \frac{A_3}{4}\|g\|^2_{\mathrm{L}^2_t} + C\mathrm{e}^{-t/2}M^2 \ , \\
A_4 \mathrm{e}^{-t/4}\|f\|_{\mathrm{H}^1_t} M &\leq \frac{A_3}{4}(\beta^2 + \|g\|^2_{\mathrm{L}^2_t}) + A_4 \mathrm{e}^{-t/4}\|f\|_{\mathrm{H}^1_t}(|\alpha| + \|f\|_{\mathrm{H}^1_t}) + C\mathrm{e}^{-t/2}M^2 \ ,
\end{aligned}$$

it follows from (3.41) that

$$\dot{S}_5 + \frac{1}{2}S_5 \leq -\frac{A_3}{2}(\beta^2 + \|g\|^2_{\mathrm{L}^2_t}) + A_4 \mathrm{e}^{-t/4}\|f\|_{\mathrm{H}^1_t}(|\alpha| + \|f\|_{\mathrm{H}^1_t}) + \tilde{C}_5 \mathrm{e}^{-t/2} M^2$$

for some $\tilde{C}_5 > 0$. Setting $\rho_0^2 = \tilde{C}_1\tilde{C}_2 A_6 \Phi_\eta(t_0, u_0, v_0)$, we have $\alpha(t)^2 + \|f(t)\|^2_{\mathrm{H}^1_t} \leq \rho_0^2$ by (3.43), (3.44), (3.45) and $\|f(t)\|^2_{\mathrm{H}^1_t} \leq A_2 S_5(t)$ by (3.40). Therefore, assuming that $\tilde{C}_5 \mathrm{e}^{-t_0/4} \leq A_3/4$, we find

$$\dot{S}_5 + \frac{1}{2}S_5 \leq -\frac{A_3}{4}(\beta^2 + \|g\|^2_{\mathrm{L}^2_t}) + \tilde{C}_6\rho_0 \mathrm{e}^{-t/4} S_5^{1/2} + \tilde{C}_5 \rho_0^2 \mathrm{e}^{-t/2} \ , \quad t \geq t_0 \ ,$$

for some $\tilde{C}_6 > 0$. Integrating this differential inequality and using the bound $S_5(t_0) \leq A_2\rho_0^2$, we obtain after a short computation

$$S_5(t) + \int_{t_0}^{t} \mathrm{e}^{-(t-s)/2}(\beta(s)^2 + \|g(s)\|_{\mathrm{L}^2_s}^2)\,\mathrm{d}s \;\leq\; \tilde{C}_7\rho_0^2(1+(t-t_0)^2)\mathrm{e}^{-(t-t_0)/2} \tag{4.1}$$

for $t \geq t_0$, where $\tilde{C}_7 > 0$ is independent of t_0 and ρ_0. In view of (3.40), this implies, in particular,

$$\|f(t)\|_{\mathrm{H}^1_t}^2 + \eta\mathrm{e}^{-t}(\beta(t)^2 + \|g(t)\|_{\mathrm{L}^2_t}^2) \;\leq\; A_2\tilde{C}_7\rho_0^2(1+(t-t_0)^2)\mathrm{e}^{-(t-t_0)/2}, \quad t \geq t_0\ . \tag{4.2}$$

Since $\dot{\alpha}(t) = \beta(t)$, we also deduce from (4.1), by a simple argument, that $\alpha(t)$ converges to some real number α^* as $t \to +\infty$ and that

$$|\alpha(t)-\alpha^*|^2 + \int_{t_0}^{t} \mathrm{e}^{-(t-s)/2}|\alpha(s)-\alpha^*|^2\,\mathrm{d}s \;\leq\; \tilde{C}_8\rho_0^2(1+(t-t_0)^2)\mathrm{e}^{-(t-t_0)/2} \tag{4.3}$$

for some $\tilde{C}_8 > 0$ and all $t \geq t_0$. Finally, it follows from (2.9) that

$$\begin{aligned}
&\|u(t) - \alpha^*\varphi^*\|_{\mathrm{H}^1_t} \leq \|f(t)\|_{\mathrm{H}^1_t} + |\alpha(t)-\alpha^*|\,\|\varphi(t)\|_{\mathrm{H}^1_t} + |\alpha^*|\,\|\varphi(t)-\varphi^*\|_{\mathrm{H}^1_t}\ ,\\
&\|v(t) - \alpha^*\psi^*\|_{\mathrm{L}^2_t} \leq \|g(t)\|_{\mathrm{L}^2_t} + |\beta(t)|\,\|\varphi\|_{\mathrm{L}^2_t} + |\alpha(t)-\alpha^*|\,\|\psi\|_{\mathrm{L}^2_t} + |\alpha^*|\,\|\psi(t)-\psi^*\|_{\mathrm{L}^2_t};
\end{aligned}$$

hence the estimate (1.23) is a direct consequence of (2.28), (2.29), (4.1), (4.2), and (4.3). This concludes the proof of Theorem 1.2. □

Acknowledgment. We thank the referees for their helpful suggestions and comments.

REFERENCES

[AW] D. G. Aronson and H. F. Weinberger, *Multidimensional nonlinear diffusion arising in population genetics*, Adv. Math., 30 (1978), pp. 33–76.

[Br] M. Bramson, *Convergence of solutions of the Kolmogorov equation to travelling waves*, Mem. Amer. Math. Soc., 44 (1983).

[BK1] J. Bricmont and A. Kupiainen, *Stability of moving fronts in the Ginzburg-Landau equation*, Comm. Math. Phys., 159 (1994), pp. 287–318.

[BK2] J. Bricmont and A. Kupiainen, *Stable non-Gaussian diffusive profiles*, Nonlinear Anal., 26 (1996), pp. 583–593.

[CH] Th. Cazenave and A. Haraux, *Introduction aux Problèmes d'Evolution Semi-linéaires*, Math. Appl. 1, Ellipses, Paris, 1990.

[DO] S. R. Dunbar and H. G. Othmer, *On a nonlinear hyperbolic equation describing transmission lines, cell movement, and branching random walks*, in Nonlinear Oscillations in Biology and Chemistry, Lecture Notes in Biomath. 66, H. G. Othmer, ed., Springer-Verlag, Berlin, 1986, pp. 274–289.

[EW] J.-P. Eckmann and C. E. Wayne, *The nonlinear stability of front solutions for parabolic partial differential equations*, Comm. Math. Phys., 161 (1994), pp. 323–334.

[EKM] M. Escobedo, O. Kavian, and H. Matano, *Large time behavior of solutions of a dissipative semi-linear heat equation*, Comm. Partial Differential Equations, 20 (1995), pp. 1427–1452.

[EZ] M. Escobedo and E. Zuazua, *Large-time behavior for convection diffusion equations in* $\mathbf{R}^N$, J. Funct. Anal., 100 (1991), pp. 119–161.

[Fi] R. A. Fisher, *The advance of advantageous genes*, Ann. Eugenics, 7 (1937), pp. 355–369.

[GV] V. A. Galaktionov and J. L. Vazquez, *Asymptotic behaviour of nonlinear parabolic equations with critical exponents. A dynamical system approach*, J. Funct. Anal., 100 (1991), pp. 435–462.

[Ga] Th. Gallay, *Local stability of critical fronts in nonlinear parabolic partial differential equations*, Nonlinearity, 7 (1994), pp. 741–764.

[GM] Th. Gallay and A. Mielke, *Diffusive mixing of stable states in the Ginzburg-Landau equation*, Comm. Math. Phys., 199 (1998), pp. 71–97.

[GR1] Th. Gallay and G. Raugel, *Stability of travelling waves for a damped hyperbolic equation*, ZAMP, 48 (1997), pp. 451–479.

[GR2] Th. Gallay and G. Raugel, *Scaling variables and asymptotic expansions in damped wave equations*, J. Differential Equations, 150 (1998), pp. 42–97.

[GR3] Th. Gallay and G. Raugel, *Stability of propagating fronts in damped hyperbolic equations*, in Partial Differential Equations: Theory and Numerical Solutions, W. Jäger, J. Nečas, O. John, K. Najzar, and J. Stará, eds., Chapman and Hall/CRC Res. Notes Math. 406, Chapman and Hall/CRC, Boca Raton, FL, 2000, pp. 130–146.

[Go] S. Goldstein, *On diffusion by discontinuous movements and the telegraph equation*, Quart. J. Mech. Appl. Math., 4 (1951), pp. 129–156.

[Ha1] K. P. Hadeler, *Hyperbolic travelling fronts*, Proc. Edinburgh Math. Soc. (2), 31 (1988), pp. 89–97.

[Ha2] K. P. Hadeler, *Reaction telegraph equations and random walk systems*, in Stochastic and Spatial Structures of Dynamical Systems, S. van Strien and S. Verduyn Lunel, eds., Royal Acad. of the Netherlands, North-Holland, Amsterdam, 1996.

[Ha3] K. P. Hadeler, *Reaction transport systems in biological modeling*, in Mathematics Inspired by Biology, Lecture Notes in Math. 1714, Springer-Verlag, New York, 1999.

[HLP] G. H. Hardy, J. E. Littlewood, and G. Polya, *Inequalities*, Cambridge University Press, Cambridge, UK, 1934 (reprinted 1964).

[HL] L. Hsiao and T.-P. Liu, *Convergence to nonlinear diffusion waves for solutions of a system of hyperbolic conservation laws with damping*, Comm. Math. Phys., 143 (1992), pp. 599–605.

[Kac] M. Kac, *A stochastic model related to the telegrapher's equation*, Rocky Mountain J. Math., 4 (1974), pp. 497–509.

[Kap] T. Kapitula, *On the stability of travelling waves in weighted L^∞ spaces*, J. Differential Equations, 112 (1994), pp. 179–215.

[Kav] O. Kavian, *Remarks on the large time behavior of a nonlinear diffusion equation*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 4 (1987), pp. 423–452.

[Ki] K. Kirchgässner, *On the nonlinear dynamics of travelling fronts*, J. Differential Equations, 96 (1992), pp. 256–278.

[KPP] A. N. Kolmogorov, I. G. Petrovskii, and N. S. Piskunov, *Etude de la diffusion avec croissance de la quantité de matière et son application à un problème biologique*, Moscow Univ. Math. Bull., 1 (1937), pp. 1–25.

[Mu] J. D. Murray, *Mathematical Biology*, 2nd ed., Biomathematics 19, Springer-Verlag, New York, 1993.

[Ni] K. Nishihara, *Convergence rates to nonlinear diffusion waves for solutions of systems of hyperbolic conservation laws with damping*, J. Differential Equations, 131 (1996), pp. 171–188.

[Pa] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Appl. Math. Sci. 44, Springer-Verlag, New York, 1983.

[RK] G. Raugel and K. Kirchgässner, *Stability of fronts for a KPP-system,* II: *The critical case*, J. Differential Equations, 146 (1998), pp. 399–456.

[Sa] D. H. Sattinger, *On the stability of waves of nonlinear parabolic systems*, Adv. Math., 22 (1976), pp. 312–355.

[Wa] C.E. Wayne, *Invariant manifolds for parabolic partial differential equations in unbounded domains*, Arch. Rational Mech. Anal., 138 (1997), pp. 279–306.

[Za] E. Zauderer, *Partial Differential Equations of Applied Mathematics*, John Wiley, New York, 1983.

SIAM J. MATH. ANAL.
Vol. 32, No. 1, pp. 30–53

RADIATION REACTION AND CENTER MANIFOLDS*

MARKUS KUNZE† AND HERBERT SPOHN‡

Abstract. We study the effective dynamics of a mechanical particle coupled to a wave field and subject to the slowly varying potential $V(\varepsilon q)$ with ε small. To lowest order in ε the motion of the particle is governed by an effective Hamiltonian. In the next order one obtains "dissipative" terms which describe the radiation reaction. We establish that this dissipative dynamics has a center manifold which is repulsive in the normal direction and which is global, in the sense that for given data and sufficiently small ε the solution stays on the center manifold forever. We prove that the solution of the full system is well approximated by the effective dissipative dynamics on its center manifold.

Key words. wave equation, adiabatic limit, effective friction

AMS subject classifications. 78A35, 35L05

PII. S0036141099351577

1. Introduction.

At the beginning of this century, in the context of the Maxwell–Lorentz equations, radiation reaction was one of the most outstanding problems in theoretical physics. It was left more or less unfinished when theoreticians turned to quantum electrodynamics. In this paper we study radiation reaction in the mathematically somewhat more accessible case of a scalar wave field. We believe that our results provide good indications on the effective dynamics for a charge coupled to the Maxwell field [14].

To explain in more detail the physical context we have to set up the model first. We consider a particle with position $q(t) \in \mathbb{R}^3$, momentum $p(t) \in \mathbb{R}^3$, and "charge" distribution ρ of total charge

$$\mathsf{e} = \int d^3x \rho(x).$$

We require that ρ be smooth, radial, and supported in a ball of radius R_ρ,

$$\text{(C)} \qquad \rho \in C_0^\infty(\mathbb{R}^3), \quad \rho(x) = \rho_r(|x|), \quad \rho(x) = 0 \quad \text{for} \quad |x| \geq R_\rho.$$

The particle is coupled to the scalar wave field $\phi(x,t)$ with the canonically conjugate momentum field $\pi(x,t)$, $x \in \mathbb{R}^3$. In addition the particle is subject to an external potential, V, whose properties will be listed below. We assume that the potential is *slowly varying* on the scale of the charge distribution, i.e., on the scale set by R_ρ. We formally introduce the dimensionless parameter ε, $\varepsilon \ll 1$, and consider the scale of potentials $V(\varepsilon q)$, $\varepsilon \to 0$. The equations of motion for the coupled system are

$$\begin{aligned} &\dot\phi(x,t) = \pi(x,t), \qquad \dot\pi(x,t) = \Delta\phi(x,t) - \rho(x - q(t)), \\ &\dot q(t) = \frac{p(t)}{\sqrt{1+p(t)^2}}, \quad \dot p(t) = -\varepsilon \nabla V(\varepsilon q(t)) + \int d^3x\, \phi(x,t) \nabla\rho(x - q(t)). \end{aligned} \tag{1.1}$$

*Received by the editors February 9, 1999; accepted for publication (in revised form) September 21, 1999; published electronically June 22, 2000.
http://www.siam.org/journals/sima/32-1/35157.html

†Mathematisches Institut der Universität Köln, Weyertal 86, D-50931 Köln, Germany (mkunze@mi.uni-koeln.de).

‡Zentrum Mathematik and Physik Department, TU München, D-80290 München, Germany (spohn@mathematik.tu-muenchen.de).

The dynamics governed by (1.1) has three distinct time scales, well separated as $\varepsilon \to 0$. On the *microscopic* time scale, $t = \mathcal{O}(1)$, the particle moves along an essentially straight line and the field adjusts itself stationarily. On a time scale $\mathcal{O}(\varepsilon^{-1})$, which we call the *macroscopic* scale, the particle feels the potential and responds to it with an effective kinetic energy which incorporates the coupling to the field. This scale was studied in [5]. The particle loses energy through radiation at a rate roughly proportional to $\ddot{q}(t)^2$. Thus on the macroscopic time scale, friction through radiation is of order ε. To resolve such an effect we have to go to an even longer time scale or must look with higher precision. The *friction* time scale is the subject of our paper.

The dynamics of (1.1) is of Hamiltonian form. We need a few facts in case the external potential vanishes, i.e., $V = 0$. Then (1.1) has the energy

$$\mathcal{H}_0(\phi, \pi, q, p) = (1 + p^2)^{1/2} + \frac{1}{2} \int d^3x \, (|\pi(x)|^2 + |\nabla\phi(x)|^2) + \int d^3x \, \phi(x)\rho(x - q)$$

and the conserved total momentum

$$\mathcal{P}(\phi, \pi, q, p) = p + \int d^3x \, \phi(x) \nabla\pi(x).$$

The minimum of $\mathcal{H}_0$, at fixed $\mathcal{P}$, is attained at

$$S_{q,v} = (\phi_v(x - q), \pi_v(x - q), q, p_v), \tag{1.2}$$

where $v \in \mathcal{V} = \{v : |v| < 1\}$, $p_v = v/\sqrt{1 - v^2}$, $\pi_v = -v \cdot \nabla\phi_v$, and $\hat{\phi}_v(k) = -\hat{\rho}(k)/[k^2 - (v \cdot k)^2]$; the hat denotes Fourier transform. We call $S_{q,v}$ the soliton centered at q, v. It has the normalized energy

$$\begin{aligned} \mathcal{E}_s(v) &= \mathcal{H}_0(S_{q,v}) - \mathcal{H}_0(S_{q,0}) \\ &= (1 - v^2)^{-1/2} - 1 + 3m_e \left[\frac{2 - v^2}{2(1 - v^2)} - \frac{1}{2|v|} \log \frac{1 + |v|}{1 - |v|} \right] \end{aligned}$$

and the total momentum

$$\begin{aligned} \mathcal{P}_s(v) &= \mathcal{P}(S_{q,v}) \\ &= v(1 - v^2)^{-1/2} + 3m_e v \left[\frac{1}{2v^2(1 - v^2)} - \frac{1}{4|v|^3} \log \frac{1 + |v|}{1 - |v|} \right]. \end{aligned} \tag{1.3}$$

Here $m_e = \frac{1}{3} \int d^3k \, |\hat{\rho}(k)|^2 k^{-2}$ is the mass of the particle due to the coupling to the field. We note that because of the Hamiltonian structure we have the identity $v(d\mathcal{P}_s/dv) = (d\mathcal{E}_s/dv)$.

Taking $S_{q,v}$ as initial conditions for (1.1) with $V = 0$ we obtain a solution travelling at constant velocity v,

$$S_{q,v}(t) = (\phi_v(x - q - vt), \pi_v(x - q - vt), q + vt, p_v), \quad v \in \mathcal{V}.$$

Let us call $\{S_{q,v} : q \in \mathbb{R}^3, v \in \mathcal{V}\}$ the six-dimensional soliton manifold, $\mathcal{S}$. Thus, for $V = 0$, if we start initially on $\mathcal{S}$ the solution remains on $\mathcal{S}$ and moves along the straight line $t \mapsto q^0 + v^0 t$. In fact, if we start close to $\mathcal{S}$, then $\mathcal{S}$ is approached asymptotically [6]. When the particle is subject to a slowly varying external potential, then the rough picture is that the solution will remain close to $\mathcal{S}$ in the course of time. For simplicity we assume throughout that the initial datum for (1.1) lies exactly on $\mathcal{S}$, i.e.,

$$(\phi(0), \pi(0), q(0), p(0)) = S_{q^0, v^0}. \tag{1.4}$$

Possible generalizations are discussed below.

At this point it is instructive to transform (1.1) to the macroscopic space-time scale in such a way that the field energy remains constant. Then the macroscopic variables, denoted by a $'$, are

$$t = \varepsilon^{-1}t', \quad x = \varepsilon^{-1}x', \quad q(t) = \varepsilon^{-1}q'(t'), \quad \text{and} \quad \phi(x,t) = \sqrt{\varepsilon}\phi'(x',t').$$

We also set

$$\rho_\varepsilon(x) = \varepsilon^{-3}\rho(\varepsilon^{-1}x).$$

In particular, $\rho_\varepsilon(x) = 0$ for $|x| \geq \varepsilon R_\rho$ and $\int d^3x \rho_\varepsilon(x) = \int d^3x \rho(x)$. With this convention, omitting the primes and indicating explicitly only the ε-dependence of $q'(t')$, we arrive at

$$\begin{aligned} \ddot{\phi}(x,t) &= \Delta\phi(x,t) - \sqrt{\varepsilon}\rho_\varepsilon(x - q^\varepsilon(t)), \\ \dot{q}^\varepsilon(t) &= v^\varepsilon(t), \\ m_0(v^\varepsilon(t))\dot{v}^\varepsilon(t) &= -\nabla V(q^\varepsilon(t)) + \sqrt{\varepsilon}\int d^3x\, \phi(x,t)\nabla\rho_\varepsilon(x - q^\varepsilon(t)). \end{aligned} \tag{1.5}$$

Here $m_0(v)$ is the 3×3 matrix defined through $m_0(v)\dot{v} = \gamma\dot{v} + \gamma^3(v \cdot \dot{v})v$ with $\gamma(v) = 1/\sqrt{1-v^2}$. Rather than momenta as in (1.1) we use velocities, which turns out to be more convenient in our context. The initial soliton (1.4) transforms to

$$S^\varepsilon_{q^0,v^0} = (\phi^\varepsilon_{v^0}(x - q^0), \pi^\varepsilon_{v^0}(x - q^0), q^0, v^0), \tag{1.6}$$

where $\hat{\phi}^\varepsilon_v(k) = -\sqrt{\varepsilon}\hat{\rho}(\varepsilon k)/[k^2 - (v \cdot k)^2]$ and $\pi^\varepsilon_v = -v \cdot \nabla\phi^\varepsilon_v$. Thus, on the macroscopic scale, the total charge is $\sqrt{\varepsilon}\int d^3x\rho(x)$, whereas

$$m_e = \frac{1}{3}\varepsilon\int d^3k |\hat{\rho}_\varepsilon(k)|^2 k^{-2}$$

is independent of ε. Equations (1.5) are again of Hamiltonian form. The energy

$$\begin{aligned} \mathcal{H}_{mac}(\phi,\pi,q,v) &= \gamma(v) + V(q) + \frac{1}{2}\int d^3x\, (|\pi(x)|^2 + |\nabla\phi(x)|^2) \\ &\quad + \sqrt{\varepsilon}\int d^3x\, \phi(x)\rho_\varepsilon(x - q) \end{aligned} \tag{1.7}$$

is conserved under (1.5). It is bounded from below, as $\mathcal{H}_{mac}(\phi,\pi,q,v) \geq V(q) - 3m_e$ independently of ε.

There is another, very instructive way to think about the initial value problem (1.5), (1.6). We prescribe initial data at $t = -\tau$, $\tau > 0$, which have finite energy and some smoothness. We refer to [6] for the precise conditions. We solve (1.5) for $V = 0$ up to time $t = 0$. Then in the limit $\tau \to \infty$ the data at $t = 0$ are exactly of the form (1.6). For $t > 0$ the external forces are acting on the particle. Clearly this causes some mismatch, which is reflected by a nonsmoothness of the fields (ϕ, π) at the light cone $\{x : |x| = t, t > 0\}$ in the limit $\varepsilon \to 0$.

Under suitable assumptions on V and for $|\rho|_{L^2}$ sufficiently small we prove in [5] that

$$|\dot{q}^\varepsilon(t)| \leq \bar{v} < 1\,, \quad |\ddot{q}^\varepsilon(t)| \leq C\,, \quad \text{and} \quad |\dddot{q}^\varepsilon(t)| \leq C \tag{1.8}$$

uniformly in ε and $t \in \mathbb{R}$ and that the limit

$$\lim_{\varepsilon \to 0^+} q^\varepsilon(t) = r(t) \tag{1.9}$$

exists. Here $r(t)$ is the solution of Hamilton's equations of motion with the effective Hamiltonian $E(p) + V(q)$ (cf. the definition of $E(p)$ below (1.3)), which in terms of velocities reads

$$\dot r = u\,, \quad m(u)\dot u = -\nabla V(r), \tag{1.10}$$

with initial data $r(0) = q^0$, $u(0) = v^0$. Here $m(u) = m_0(u) + m_f(u)$, where $m_f(u)$ is the additional "mass" due to the coupling to the field defined by

$$m_f(u)\dot u = 3m_e \left(\varphi(|u|)\dot u + |u|^{-1}\varphi'(|u|)(u\cdot \dot u)u\right) \tag{1.11}$$

as a 3×3 matrix, where $\varphi(|v|)$ is the function appearing in the square brackets of (1.3). Note that the energy

$$H(r,u) = \mathcal{E}_s(u) + V(r) \tag{1.12}$$

is conserved by the solutions to (1.10).

With this background information let us return to the radiation reaction as discussed by Abraham, Lorentz, Schott, and Dirac; cf. [16] for an excellent account. Of course, these theoretical physicists were interested in the electrodynamics of moving charges. We take here the liberty to transcribe their arguments to the case of a scalar wave equation. For the sake of discussion, we reintroduce the bare mass m_0 and state the equations for small velocities only. In our proof below, however, we will handle all $v \in \mathcal{V}$.

At the beginning of this century, the hope was to define a structureless elementary charge through a point charge limit. For this program, one had to model the charge distribution phenomenologically with the understanding that finer details should become irrelevant in the limit. In (1.1) we adopted the Abraham model of a rigid charge distribution. The point charge limit then corresponds to taking in (1.1) a fixed ε-independent potential and to let the diameter of the charge distribution tend to zero. If this diameter is set proportional to ε and if we compare with (1.5), then in the point charge limit the charge distribution $\sqrt{\varepsilon}\rho_\varepsilon$ is to be replaced by ρ_ε, which in particular shows that the adiabatic limit of a slowly varying external potential is distinct from the point charge limit, where $\mathsf{e} = \int d^3x \rho_\varepsilon(x) = \int d^3x \rho(x)$ is independent of ε and the electromagnetic mass diverges as $\frac{1}{3}\int d^3k |\hat\rho_\varepsilon(k)|^2 k^{-2} = \varepsilon^{-1} m_e$. A formal Taylor expansion leads to the effective equation of motion

$$m_0 \ddot r = -\nabla V(r) - \varepsilon^{-1} m_e \ddot r + a\mathsf{e}^2\, \dddot r, \tag{1.13}$$

valid for small velocities $\dot r$, with some constant $a > 0$. Equation (1.13) is the nonrelativistic limit of the Lorentz–Dirac equation [10]. The standard argument, reproduced in many textbooks, such as [3] (with the notable exception of Landau and Lifshitz [9]), is to lump m_0 and $\varepsilon^{-1} m_e$ together and to take the limits $\varepsilon \to 0$ and $m_0 \to -\infty$ at constant $m_0 + \varepsilon^{-1} m_e = m_{exp}$, the experimentally observed mass of the particle. Then (1.13) reads as

$$m_{exp} \ddot r = -\nabla V(r) + a\mathsf{e}^2\, \dddot r\,. \tag{1.14}$$

Since this equation is of third order, one needs besides q^0, v^0 also $\dot{u}(0)$ as an initial condition which has to be extracted somehow from the initial data of the full system. Even worse, (1.14) has solutions which are exponentially unbounded in time, the famous run-away solutions. Thus one needs an additional criterion to single out the solutions of physical relevance. Dirac [1], and later Haag [2], argued that physical solutions have to satisfy the asymptotic condition

$$\lim_{t\to\infty} \ddot{r}(t) = 0, \tag{1.15}$$

as a substitute for the missing initial condition $\ddot{r}(0)$. The validity of the asymptotic condition has been checked only in trivial cases; see [10]. For general V one should expect the solutions to (1.13) to be chaotic. Physical and unphysical solutions might become badly mixed up. On a more practical level, the physical solutions are unstable and therefore difficult to compute numerically. To put it in the words of Thirring [15]: "...(1.14) *has not only crazy solutions and there are attempts to separate sense from nonsense through special initial conditions. But one hopes that the true solution to the problem will look differently and that the nature of the equations of motion is not so highly unstable that the act of balance can be achieved only through a stroke of good fortune in the initial conditions.*"

This is indeed the case, as we are going to show in this paper, and our resolution requires just a little twist. If instead of the point charge limit we consider a slowly varying external potential, then on the macroscopic time scale, according to (1.5), equation (1.13) reads

$$(m_0 + m_e)\ddot{r} = -\nabla V(r) + \varepsilon a \mathsf{e}^2 \dddot{r}, \tag{1.16}$$

which reflects that radiation reaction is a small correction to the Hamiltonian motion. In (1.16) the highest derivative appears with a small prefactor. Such differential equations are studied in geometric singular perturbation theory. From there we know that (1.16) has a six-dimensional invariant center manifold $\mathcal{I}_\varepsilon$, which is only $\mathcal{O}(\varepsilon)$ away from the Hamiltonian manifold $\mathcal{I}_0 = \{(q, \dot{q}, \ddot{q}) : (m_0 + m_e)\ddot{q} = -\nabla V(q)\}$. For initial conditions slightly off $\mathcal{I}_\varepsilon$ the solution moves away from $\mathcal{I}_\varepsilon$ exponentially fast. On $\mathcal{I}_\varepsilon$, $\dot{q}$ is bounded away from 1, $\ddot{q}$ is bounded, and the motion is governed by an effective second order equation (cf. (4.9) below), which gives *precisely* the physical solutions. To establish such a result we have to prove that the solution to (1.5) indeed stays close to $\mathcal{I}_\varepsilon$.

In our paper we carry out this program, essentially under the same conditions as in [5], namely, a sufficiently differentiable V and $|\rho|_{L^2}$ small. Our main additional estimate is

$$|\dddot{v}^{\,\varepsilon}(t)| \leq C \tag{1.17}$$

uniformly in ε and $t \in \mathbb{R}$. Thereby we can bound one further order in the rigorous Taylor expansion and obtain, setting $\dot{q}^\varepsilon = v^\varepsilon$,

$$m(v^\varepsilon)\dot{v}^\varepsilon = -\nabla V(q^\varepsilon) + \varepsilon a(v^\varepsilon)\ddot{v}^\varepsilon + \varepsilon b(v^\varepsilon, \dot{v}^\varepsilon) + \varepsilon^2 f^\varepsilon(t), \quad t \geq \varepsilon t_1, \tag{1.18}$$

with $|f^\varepsilon(t)| \leq C$ and coefficient functions a, b that will be defined below. Clearly (1.18) should be compared with

$$\dot{r} = u\,, \quad m(u)\dot{u} = -\nabla V(r) + \varepsilon a(u)\ddot{u} + \varepsilon b(u, \dot{u}). \tag{1.19}$$

Our crucial observation is that the condition $|u(t)| \leq \text{const.} < 1$ for all t holds only on the center manifold $\mathcal{I}_\varepsilon$. Thus the a priori estimate $|\dot{q}^\varepsilon(t)| \leq \bar{v} < 1$ (see (1.8)), together with the initial conditions $r(0) = q^0$, $u(0) = v^0$, *uniquely* singles out that solution of (1.19) which is to be compared with the true solution.

The coefficient functions a, b are proportional to e^2. If the total charge $\mathsf{e} = 0$, then the friction term in (1.18) vanishes identically and radiation reaction appears at a higher level of approximation.

Since on the error term $f^\varepsilon(t)$ in (1.18) we know only that it is uniformly bounded, the difference $|q^\varepsilon(t) - r(t)|$, with $r(t)$ having initial conditions on $\mathcal{I}_\varepsilon$, can be bounded at best as εe^{ct}. Thus on the time scale $t = \mathcal{O}(1)$ we seem to be back to the result (1.9) already proved in [5]. To distinguish, from this point of view, between (1.19) and (1.10) we would have to control the difference with a precision of order ε^2. At present we do not know whether this is possible, but nevertheless we can prove the weaker statement

$$|H(q^\varepsilon(t), v^\varepsilon(t)) - H(r(t), u(t))| \leq \text{const.}\, \varepsilon^2, \tag{1.20}$$

where H is the energy from (1.12). Thus, on a surface of constant energy the difference $|q^\varepsilon(t) - r(t)|$ could be of order ε, whereas along ∇H it must be of order ε^2. In addition to (1.20) it may also be shown that in fact $|q^\varepsilon(t) - r(t)| \sim \varepsilon^3$ on the short time scale $t = \mathcal{O}(\varepsilon)$, a result that is quite natural from the viewpoint of singularly perturbed ODEs. On the original time scale of (1.1) this amounts at least to an estimate with precision ε^2 over time intervals of length $\mathcal{O}(1)$, a result that could not have been obtained from the bounds in [5].

Taking a somewhat broader perspective, the problem discussed here may be viewed as an infinite-dimensional Hamiltonian system which relaxes to a stationary solution through the emission of radiation. This phenomenon is fairly common and has been studied in the context of linear and nonlinear wave equations. We refer to [12, 13] for recent work, which also contain more references to prior studies. For such problems one typically has a stationary eigenmode which turns into a resonance by coupling to propagating modes. In comparison, a simplifying feature in the present paper is that the localized and propagating degrees of freedom are already well separated on the level of the equations of motion. On the other hand, we provide a quantitative estimate on the relaxation process and not just a power law decay in time.

2. Main results. We give some more details and state our main results precisely. First we have to establish the bound (1.17).

LEMMA 2.1. *For $|\rho|_{L^2}$ sufficiently small we have*

$$\sup_{t \in \mathbb{R}} |\dddot{v}^\varepsilon(t)| \leq C$$

for every solution of (1.5) *which starts on the soliton manifold $\mathcal{S}$. Both the constant C and the bound for $|\rho|_{L^2}$ depend only on the initial data.*

The bound of Lemma 2.1 may be used to Taylor expand the self-force

$$F_s^\varepsilon(t) = \sqrt{\varepsilon} \int d^3x\, \phi(x,t) \nabla \rho_\varepsilon(x - q^\varepsilon(t)) \tag{2.1}$$

in (1.5) as

$$F_s^\varepsilon(t) = -m_f(v^\varepsilon(t))\dot{v}^\varepsilon(t) + \varepsilon a(v^\varepsilon(t))\ddot{v}^\varepsilon(t) + \varepsilon b(v^\varepsilon(t), \dot{v}^\varepsilon(t)) + \mathcal{O}(\varepsilon^2)\,, \quad t \geq \varepsilon t_1\,, \tag{2.2}$$

which together with the second equation in (1.5) yields (1.18). Here m_f is defined in (1.11). $t_1 = 2R_\rho/(1-\bar{v})$ is the microscopic time for the wave equation to forget its initial data through the compact support of ρ and the velocity bound; cf. assumption (C) and (1.8). The coefficient functions are given by

$$a(v)\ddot{v} = (\mathsf{e}^2/24\pi)(\ddot{v}\cdot\nabla_v)\nabla_v\gamma^2 = (\mathsf{e}^2/12\pi)[\gamma^4\ddot{v} + 4\gamma^6(v\cdot\ddot{v})v]\,, \tag{2.3}$$

$$\begin{aligned} b(v,\dot{v}) &= (\mathsf{e}^2/32\pi)(\dot{v}\cdot\nabla_v)^2\nabla_v\gamma^2 \\ &= (\mathsf{e}^2/4\pi)[2\gamma^6(v\cdot\dot{v})\dot{v} + \gamma^6\dot{v}^2v + 6\gamma^8(v\cdot\dot{v})^2v], \end{aligned} \tag{2.4}$$

$\dot{v}, \ddot{v} \in \mathbb{R}^3$, with $\gamma = 1/\sqrt{1-v^2}$, $|v| < 1$.

Next we explain the existence and the role of the center-like manifolds $\mathcal{I}_\varepsilon$ in greater detail. We refer to [11, 4] for further background on geometric singular perturbation theory. To rewrite (1.19) as a singular perturbation problem, let

$$x = (r,u) \in \mathbb{R}^3\times\mathcal{V}\,, \quad y = \dot{u} \in \mathbb{R}^3\,, \quad f(x,y) = (x_2,y) \in \mathcal{V}\times\mathbb{R}^3\,, \quad \text{and}$$
$$g(x,y,\varepsilon) = a(x_2)^{-1}[m(x_2)y + \nabla V(x_1) - \varepsilon b(x_2,y)]\,.$$

Then (1.19) reads as

$$\dot{x} = f(x,y)\,, \quad \varepsilon\dot{y} = g(x,y,\varepsilon)\,. \tag{2.5}$$

We intend to apply the results from [11] to (2.5) in order to find a center-like manifold for the perturbed problem near the corresponding manifold for the $(\varepsilon = 0)$-problem. With $h(x) = -m(x_2)^{-1}\nabla V(x_1)$, let

$$\begin{aligned} \mathcal{I}_0 &= \{(x,y) : g(x,y,0) = 0\} = \{(r,u,\dot{u}) : m(u)\dot{u} = -\nabla V(r)\} \\ &= \{(x,h(x)) : x \in \mathbb{R}^3\times\mathcal{V}\} \end{aligned} \tag{2.6}$$

be this invariant manifold for (2.5) with $\varepsilon = 0$. The flow on $\mathcal{I}_0$ is governed by the equation $\dot{x} = f(x,h(x))$, or stated differently, $m(\dot{r})\ddot{r} = -\nabla V(r)$, the familiar Hamiltonian flow.

To see that $\mathcal{I}_0$ is perturbed to some $\mathcal{I}_\varepsilon$ with ε small, we have to modify the functions $a(u)$, $m(u)$, and $b(u,\dot{u})$ for $|u|$ close to 1 due to the singularity at $|u| = 1$. This will cause no problems later on, since we already have the a priori bound $|v^\varepsilon(t)| \le \bar{v} < 1$ for the velocity of the true system. In (4.4) below, we will fix a small $\bar{\delta} = \bar{\delta}(\bar{v}) > 0$ satisfying some estimates; $\bar{\delta}$ depends only on bounds for the initial data, since $\bar{v}$ does so. Let

$$\mathcal{K}_{1-\bar{\delta}} = \mathbb{R}^3\times\{u\in\mathbb{R}^3 : |u| \le 1-\bar{\delta}\}.$$

We continue $a(u)$, $m(u)$, and $b(u,\dot{u})$ with their values at $|u| = 1-\bar{\delta}$ to the missing infinite strip $1-\bar{\delta} < |u| < 1$. Then the basic assumptions (I), (II) from [11, p. 45] are satisfied, since $\mathcal{I}_0$ is also what is called normally hyperbolic, i.e., repulsive in the direction normal to $\mathcal{I}_0$ at an ε-independent rate; see Lemma 4.1 below. Hence we find $\varepsilon_0 = \varepsilon_0(\bar{\delta}) > 0$ and a C^1-function $h(x,\varepsilon) = h_\varepsilon(x) : \mathbb{R}^3\times\mathcal{V}\times]0,\varepsilon_0] \to \mathbb{R}^3$ such that for $\varepsilon \le \varepsilon_0$,

$$\mathcal{I}_\varepsilon = \{(x,h_\varepsilon(x)) : x\in\mathbb{R}^3\times\mathcal{V}\}$$

is forward invariant for the flow (1.19) with the modified functions a, m, b. Since the modified equation agrees with (1.19) in the interior of $\mathcal{K}_{1-\bar{\delta}}$, we conclude that $\mathcal{I}_\varepsilon$

is locally invariant for the flow (1.19); i.e., the solution of the modified equation is the solution to the original equation as long as it does not reach the boundary set $\{(x, h_\varepsilon(x)) = (r, u, h_\varepsilon(r,u)) : |u| = 1 - \bar{\delta}\}$. The flow for $\varepsilon = 0$ is then perturbed to $\dot{x} = f(x, h_\varepsilon(x))$ for $\varepsilon \le \varepsilon_0$.

We will show in Theorem 4.4 below that for $\varepsilon \in]0, \varepsilon_1]$, with $\varepsilon_1 > 0$ sufficiently small, all solutions of (1.19) starting at points $(r, u, h_\varepsilon(r,u)) \in \mathcal{I}_\varepsilon$ with $|u| \le \bar{v}$ will indeed stay away from the boundary $\{(r, u, h_\varepsilon(r,u)) : |u| = 1 - \bar{\delta}\}$ for all future times. In addition, $\nabla V(r(t)) \to 0$ and $\ddot{r}(t) \to 0$ as $t \to \infty$, which is just the asymptotic condition (1.15) postulated by Dirac [1] and Haag [2]. If the potential is sufficiently confining, then the solution trajectory on $\mathcal{I}_\varepsilon$ not only approaches the set of critical points for V in the long time limit, but it converges to some definite critical point. Moreover, we will show that for all solutions on the center manifold, $\dot{u}(t)$ and $\ddot{u}(t)$ are bounded, and $u(t)$ is bounded away from 1, uniformly in ε and t. Conversely, every such solution to (1.19) has to lie on $\mathcal{I}_\varepsilon$. Thus $\mathcal{I}_\varepsilon$ indeed characterizes the physical solutions.

To summarize, we have established now the existence of a center manifold $\mathcal{I}_\varepsilon$ with a well-defined (semi-) flow on it that gives a unique solution to (1.19) for initial velocities bounded by $\bar{v}$.

For the potential $V \in C^3(\mathbb{R}^3)$ we assume that it is bounded in the sense that

$$\text{(U)} \qquad \sup_{q \in \mathbb{R}^3} \Big(|V(q)| + |\nabla V(q)| + |\nabla\nabla V(q)| + |\nabla\nabla\nabla V(q)| \Big) < \infty \,.$$

The method works equally well for $V \in C^3(\mathbb{R}^3)$ which is confining, i.e.,

$$\text{(U}') \qquad V(q) \to \infty \quad \text{as} \quad |q| \to \infty \,,$$

as will be made more precise in section 4; cf. Theorem 4.8.

Our main result is the following.

THEOREM 2.2. *Assume* (U) *or* (U′) *for the potential, and let the initial data* $(\phi^0(x), \pi^0(x), q^0, v^0)$ *for* (1.5) *be given by* (1.6). *Let* $|\rho|_{L^2}$ *and* $\varepsilon \le \varepsilon_1$ *be sufficiently small, and introduce the center manifolds* $\mathcal{I}_\varepsilon$ *for the comparison dynamics* (1.19) *as explained above. At time* $\varepsilon t_1 = \varepsilon 2R_\rho/(1-\bar{v})$ *we match the initial values* $r(\varepsilon t_1) = q^\varepsilon(\varepsilon t_1)$, $u(\varepsilon t_1) = v^\varepsilon(\varepsilon t_1)$ *for the motion on the center manifold; i.e., the initial data for the comparison dynamics are*

$$(q^\varepsilon(\varepsilon t_1), v^\varepsilon(\varepsilon t_1), h_\varepsilon(q^\varepsilon(\varepsilon t_1), v^\varepsilon(\varepsilon t_1))) \in \mathcal{I}_\varepsilon \,.$$

Then for every $\tau > 0$ *there exists* $c(\tau) > 0$ *such that for all* $t \in [\varepsilon t_1, \varepsilon t_1 + \tau]$

$$(2.7) \quad |q^\varepsilon(t) - r(t)| \le c(\tau)\varepsilon \,, \quad |v^\varepsilon(t) - u(t)| \le c(\tau)\varepsilon \,, \quad \textit{and} \quad |\dot{v}^\varepsilon(t) - \dot{u}(t)| \le c(\tau)\varepsilon.$$

In addition we have the bound

$$(2.8) \qquad |H(q^\varepsilon(t), v^\varepsilon(t)) - H(r(t), u(t))| \le c(\tau)\varepsilon^2.$$

Remark 2.3. (i) As already mentioned at the end of the introduction, we can also show

$$(2.9) \qquad |q^\varepsilon(t) - r(t)| \le c(\tau)\varepsilon^3 \quad \text{and} \quad |v^\varepsilon(t) - u(t)| \le c(\tau)\varepsilon^2$$

for $t \in [\varepsilon t_1, \varepsilon t_1 + \varepsilon\tau]$, i.e., $t = \mathcal{O}(\varepsilon)$; cf. Proposition 5.1.

(ii) The construction of the center manifolds and the upper bound for $|\rho|_{L^2}$ rely only on bounds for the data, but not on properties of a particularly chosen solution. Our main technical assumption is a sufficiently small $|\rho|_{L^2}$ which is presumably not necessary.

(iii) In [5] we did not require the true solution to start on the soliton manifold, but instead to start close to it. We refer to the criterion [5, Thm. 2.6] for an "adiabatic" family of solutions. The same generality could be achieved in the present context, using an appropriately modified version of [5, Thm. 2.6]. In section 8 we derive the relevant estimates, in particular (8.8), in full generality containing a nonzero initial difference $Z(0)$. The corresponding generalization of Theorem 2.2 is then straightforward. However, since we did not want to obscure our main achievement through technicalities, we decided to elaborate here the more accessible case of a trajectory starting right on the soliton manifold. In the same spirit we do not consider arbitrary time intervals of length τ, but only the particular $[\varepsilon t_1, \varepsilon t_1 + \tau]$.

(iv) The existence of solutions to (1.1) is discussed in [5, Lem. 2.2]. For every initial value $Y^0 = (\phi^0(x), \pi^0(x), q^0, p^0) \in \mathcal{E}$ we find a unique (weak) solution $Y(\cdot) \in C(\mathbb{R}, \mathcal{E})$ such that $Y(0) = Y^0$. Here the state space is $\mathcal{E} = D^{1,2}(\mathbb{R}^3) \oplus L^2(\mathbb{R}^3) \oplus \mathbb{R}^3 \oplus \mathbb{R}^3$ (where $D^{1,2}(\mathbb{R}^3) = \{\phi \in L^6(\mathbb{R}^3) : |\nabla\phi| \in L^2(\mathbb{R}^3)\}$) with norm $|Y|_{\mathcal{E}} = |\nabla\phi|_{L^2} + |\pi|_{L^2} + |q| + |p|$.

Having such fairly precise information on the particle trajectory, we can also determine the adiabatic limit $\varepsilon \to 0$ of the fields (ϕ, π) in (1.5) through the solution of the inhomogeneous wave equation. We generate the initial data as explained in the introduction. On the level of the comparison dynamics this means to extend $r(t)$ and $u(t)$ to negative times $t \leq 0$ by $r(t) = q^0 + tv^0$, resp., $u(t) = v^0$. Let the retarded time t_{ret}, depending on x and t, be the unique solution of $t_{\mathrm{ret}} = t - |x - r(t_{\mathrm{ret}})|$, and let $\hat{n}(x,t) = (x - r(t_{\mathrm{ret}}))/|x - r(t_{\mathrm{ret}})|$.

THEOREM 2.4. *Under the conditions of Theorem* 2.2 *and for the fields* (ϕ, π) *from* (1.5) *we have for* $x \neq r(t)$ *the pointwise limits*

$$\lim_{\varepsilon\to 0} \frac{1}{\sqrt{\varepsilon}}\, \phi(x,t) = -\frac{\mathsf{e}}{4\pi|x - r(t_{\mathrm{ret}})|}\, (1 - \hat{n}(x,t)\cdot u(t_{\mathrm{ret}}))^{-1} \tag{2.10}$$

and, except for the light cone $\{x : |x| = t > 0\}$,

$$\begin{aligned} &\lim_{\varepsilon\to 0} \frac{1}{\sqrt{\varepsilon}}\, \pi(x,t) \\ &= -\frac{\mathsf{e}}{4\pi|x - r(t_{\mathrm{ret}})|}\, (1 - \hat{n}(x,t)\cdot u(t_{\mathrm{ret}}))^{-3}\, \hat{n}(x,t)\cdot \dot{u}(t_{\mathrm{ret}}) \\ &\quad -\frac{\mathsf{e}}{4\pi|x - r(t_{\mathrm{ret}})|^2}\, (1 - \hat{n}(x,t)\cdot u(t_{\mathrm{ret}}))^{-3}\, (\hat{n}(x,t)\cdot u(t_{\mathrm{ret}}) - u(t_{\mathrm{ret}})^2). \end{aligned} \tag{2.11}$$

The paper is organized as follows. Since the proof of Lemma 2.1 is rather technical, we moved it to an appendix, section 8. The derivation of the representation (2.2) of the self-force term is the contents of section 3. In section 4 we give supplementary remarks on the behavior of solutions on the center manifold, whereas in section 5 we carry out the proofs of Theorem 2.2 and Proposition 5.1. Section 6 contains the proof of Theorem 2.4, and finally in section 7 we determine the amount of energy radiated to infinity.

3. Representation of the self-force. In this section we show that the self-force $F_s^\varepsilon(t)$ from (2.1) can be written in the form (2.2). We carry out this computation

on the original fast time scale corresponding to (1.1) since we will need some of the arguments from [5]. Thus we consider

$$F_s(t) = \int d^3x\, \phi(x,t)\nabla\rho(x-q(t)).$$

Since $\phi(x,t) = \phi_r(x,t) + \phi_0(x,t)$, where $\ddot{\phi}_0 = \Delta\phi_0$ with the initial values $\phi_0(x,0) = \phi^0(x)$ and $\pi_0(x,0) = \pi^0(x)$, and since

$$\phi_r(x,t) = -\frac{1}{4\pi}\int_0^t \frac{ds}{t-s}\int_{|y-x|=t-s} d^2y\,\rho(y-q(s))$$

is the retarded potential, we can decompose accordingly,

$$F_s(t) = F_0(t) + F_r(t) = \langle\phi_0(\cdot,t), \nabla\rho(\cdot - q(t))\rangle + \langle\phi_r(\cdot,t), \nabla\rho(\cdot - q(t))\rangle\,.$$

LEMMA 3.1. *The function $F_0(t)$ vanishes for $t \geq t_1 = 2R_\rho/(1-\bar{v})$.*

Proof. Let $U(t)$ denote the group generated by the free wave equation in $D^{1,2}(\mathbb{R}^3)\oplus L^2(\mathbb{R}^3)$. Then (1.4) and Fourier transformation implies

$$(\phi^0(x), \pi^0(x)) = -\int_{-\infty}^0 ds\,[U(-s)\bar{\rho}(\cdot - q^0 - v^0 s)](x)$$

with $\bar{\rho}(x) = (0, \rho(x))$. Thus Kirchhoff's formula yields, as a consequence of $|v^0| < 1$, that $\phi_0(x,t) = 0$ for $|x - q^0| \leq t - R_\rho$. Since $|q(t) - q^0| \leq \bar{v}t$, the claim follows. ∎

Hence to show (2.2) it is enough to prove the following.

LEMMA 3.2. *For $t \geq t_1$,*

$$F_r(t) = -m_f(v(t))\dot{v}(t) + a(v(t))\ddot{v}(t) + b(v(t), \dot{v}(t)) + \mathcal{O}(\varepsilon^3);$$

cf. (1.11), (2.3), *and* (2.4).

Proof. We follow the proof of [5, Lem. 5.1] but expand

$$q(s) = q(t) - v(t)(t-s) + \frac{1}{2}\dot{v}(t)(t-s)^2 - \frac{1}{6}\ddot{v}(t)(t-s)^3 + \mathcal{O}(\varepsilon^3)$$

up to third order, which is allowed by Lemma 2.1. Through Fourier transformation we arrive at

$$F_r(t) = (-i)\int_0^t ds \int d^3k\,|\hat{\rho}(k)|^2 \frac{k}{|k|}\,\sin|k|(t-s)\,e^{-i(k\cdot v)(t-s)} \times e^{-i[-\frac{1}{2}(k\cdot\dot{v})(t-s)^2 + \frac{1}{6}(k\cdot\ddot{v})(t-s)^3]} + \mathcal{O}(\varepsilon^3)\,,$$

with $v = v(t)$, etc. As in [5, Lem. 5.1], here and in the following $\int_0^t ds(\ldots)$ can be changed forth and back to $\int_{t-T}^t ds(\ldots)$ for all $t, T \geq t_1$. Because

$$e^{-i[-\frac{1}{2}(k\cdot\dot{v})(t-s)^2 + \frac{1}{6}(k\cdot\ddot{v})(t-s)^3]} = 1 + \frac{i}{2}(k\cdot\dot{v})(t-s)^2 - \frac{i}{6}(k\cdot\ddot{v})(t-s)^3 - \frac{1}{8}(k\cdot\dot{v})^2(t-s)^4 + \mathcal{O}(\varepsilon^3)$$

for $t - s = \mathcal{O}(1)$ by (8.1) below, we obtain, for $t, T \geq t_1$,

$$F_r(t) = (-i)\int_0^T d\tau \int d^3k\,|\hat{\rho}(k)|^2 \frac{k}{|k|}\,\sin|k|\tau\,e^{-i(k\cdot v)\tau} \times \left[1 + \frac{i}{2}(k\cdot\dot{v})\tau^2 - \frac{i}{6}(k\cdot\ddot{v})\tau^3 - \frac{1}{8}(k\cdot\dot{v})^2\tau^4\right] + \mathcal{O}(\varepsilon^3)\,.$$

Let

$$I_p = \int_0^T d\tau \, \frac{\sin|k|\tau}{|k|} \, e^{-i(k\cdot v)\tau} \tau^p \,, \quad p = 0, \dots, 4 \,.$$

Then

$$(\ddot{v}\cdot\nabla_v)\nabla_v I_1 = -k(k\cdot\ddot{v})I_3 \quad \text{and} \quad (\dot{v}\cdot\nabla_v)^2\nabla_v I_1 = ik(k\cdot\dot{v})^2 I_4 \,.$$

Our claim now follows from Lemma 3.3 shown below, $\int d^3k \, |\hat{\rho}(k)|^2 k I_0 \to 0$, and $(1/2)\int d^3k \, |\hat{\rho}(k)|^2 k(k\cdot\dot{v})I_2 \to -m_f(v(t))\dot{v}(t)$ for $T \to \infty$; see [5, Appendix A]. □

LEMMA 3.3. *We have the identity*

$$\int_0^\infty dt \, t \int d^3k |\hat{\rho}(k)|^2 \frac{\sin|k|t}{|k|} \, e^{-i(k\cdot v)t} = (\mathsf{e}^2/4\pi)\gamma^2 \,.$$

Proof. Since $\hat{\rho}(k) = \hat{\rho}_r(|k|)$ is radial, and by transformation to polar coordinates,

$$\int d^3k |\hat{\rho}(k)|^2 \frac{\sin|k|t}{|k|} \, e^{-i(k\cdot v)t} = \frac{4\pi}{t|v|} \int_0^\infty dR \, |\hat{\rho}_r(R)|^2 \sin(Rt)\sin(Rt|v|) \,.$$

Thus for fixed $T > 0$,

$$\begin{aligned} &\int_0^T dt \, t \int d^3k |\hat{\rho}(k)|^2 \frac{\sin|k|t}{|k|} \, e^{-i(k\cdot v)t} \\ &= \frac{2\pi}{|v|} \int_0^\infty \frac{dR}{R} |\hat{\rho}_r(R)|^2 \left(\frac{\sin(R(1-|v|)T)}{1-|v|} - \frac{\sin(R(1+|v|)T)}{1+|v|} \right) . \end{aligned}$$

To complete the proof we need only to verify that $\int d^3k |\hat{\rho}(k)|^2 \, |k|^{-3} \sin|k|T \to \mathsf{e}^2/4\pi$ as $T \to \infty$. To see this, let $\hat{\psi}(k) = |k|^{-3}\sin(|k|T)$. Then

$$\int d^3k |\hat{\rho}(k)|^2 \, \hat{\psi}(k) = (2\pi)^{-3/2} \int d^3x \rho(x) \int d^3y \, \rho(y)\psi(x-y),$$

and we are going to show $\psi(x) \to \sqrt{\pi/2}$ as $T \to \infty$. We have, by transformation to polar coordinates,

$$(2\pi)^{3/2}\psi(x) = \int d^3k \, \hat{\psi}(k) e^{-ik\cdot x} = 4\pi \int_0^\infty ds \, \frac{\sin(s)}{s} \, \frac{\sin(s|x|/T)}{s|x|/T} \to 2\pi^2$$

for $T \to \infty$. This completes the proof. □

4. More about the center manifold. In this section we explain the behavior of solutions on the center manifold. First we show that the unperturbed manifold $\mathcal{I}_0$ from (2.6) is hyperbolic in normal direction.

LEMMA 4.1. *The eigenvalues of $D_y g(x,y,0) = a(x_2)^{-1}m(x_2)$ are bounded below by a positive constant, uniformly in $x = (r,u)$ with $r \in \mathbb{R}^3$ and $|u| \le 1-\delta$ for all prescribed $\delta \in]0,1]$.*

Proof. By [8, Thm. 2, p. 185], $a(u)$ and $m(u)$ can be simultaneously transformed to diagonal form through a single nonsingular matrix B. In addition, denoting by $b_j \neq 0$ the jth column of B and by λ_j the jth eigenvalue of $a(u)^{-1}m(u)$, one has $\lambda_j a(u)b_j = m(u)b_j$, $j = 1,2,3$. Multiplication by b_j leads to $\lambda_j(\mathsf{e}^2/12\pi)\gamma^3[\gamma b_j^2 + 4\gamma^3(v\cdot b_j)^2] \ge \gamma b_j^2 + \gamma^3(v\cdot b_j)^2$, and thus $\lambda_j \ge (3\pi/\mathsf{e}^2)\gamma^{-3}$. □

Since $a(u)$, $m(u)$ are modified to be constant outside $|u| \le 1-\bar\delta$, their corresponding eigenvalues are uniformly bounded below for $|u|<1$. As a consequence of Lemma 4.1 the manifolds $\mathcal{I}_\varepsilon$ are unstable at some exponential rate $e^{\mu t}$ for solutions in the normal direction.

We note that, by [11, Thm. 2.1],

$$\sup\{|h_\varepsilon(r,u)| : (r,u)\in \mathbb{R}^3\times\mathcal{V}, \varepsilon\in]0,\varepsilon_0]\} \le c = c(\bar\delta). \tag{4.1}$$

Our next aim is to prove global existence of solutions to (1.19) forward in time which start over $\mathcal{K}_{\bar v} = \mathbb{R}^3\times\{u\in\mathbb{R}^3 : |u|\le \bar v\}$ on the center manifold, provided $\varepsilon\le\varepsilon_1$ with $\varepsilon_1>0$ sufficiently small. For this purpose we introduce a suitable Lyapunov function.

LEMMA 4.2. *Let*

$$G_\varepsilon(r,u,\dot u) = H(r,u) - \varepsilon(a(u)\dot u)\cdot u = \mathcal{E}_s(u) + V(r) - \varepsilon(a(u)\dot u)\cdot u\,.$$

Then along solutions $(r(t),u(t),\dot u(t))$ *of* (1.19) *we have*

$$\frac{d}{dt}G_\varepsilon(r,u,\dot u) = -\varepsilon(\mathsf{e}^2/12\pi)\,[6\gamma^8(u\cdot\dot u)^2 + \gamma^6\dot u^2]. \tag{4.2}$$

Proof. Observing that

$$(a(u)\dot u)\cdot u = (\mathsf{e}^2/12\pi)\gamma^6\,(1+3u^2)(u\cdot\dot u),$$

this is a straightforward calculation. □

Through the Lyapunov function G_ε we can control the long time behavior.

THEOREM 4.3. *Let* (U) *or* (U′) *hold and let any global solution* $(r(t),u(t))$ *of* (1.19) *be given such that* $\sup_{t\ge0}|u(t)|\le\bar u(\varepsilon)<1$ *and* $\sup_{t\ge0}|\dot u(t)|\le c(\varepsilon)$, *for possibly* ε*-dependent constants* $\bar u(\varepsilon)$ *and* $c(\varepsilon)$. *Then*

$$\dot u(t)\to 0,\quad \ddot u(t)\to 0,\quad \textit{and}\quad \nabla V(r(t))\to 0\quad \textit{as}\quad t\to\infty.$$

Proof. Denoting by $c(\varepsilon)$ or $C(\varepsilon)$ general ε-dependent constants, by Lemma 4.2 we have along a trajectory

$$\begin{aligned} c(\varepsilon)\int_0^T \dot u^2\,dt &\le -\int_0^T \frac{d}{dt}G_\varepsilon\,dt \\ &= -\mathcal{E}_s(u(T)) - V(r(T)) + \varepsilon(a(u(T))\dot u(T))\cdot u(T) \\ &\quad +\mathcal{E}_s(u^0) + V(r^0) - \varepsilon(a(u^0)\dot u^0)\cdot u^0 \\ &\le C(\varepsilon,\mathrm{data}). \end{aligned}$$

For the last estimate observe $\inf_{r\in\mathbb{R}^3}V(r)>-\infty$ in both cases (U) and (U′). Thus $\int_0^\infty \dot u^2\,dt \le C(\varepsilon,\mathrm{data})$ and, by (1.19), also $\sup_{t\ge0}|\ddot u(t)|\le C(\varepsilon,\mathrm{data})$. Hence we conclude $\dot u(t)\to0$ as $t\to\infty$. Next, differentiation of (1.19) yields $\sup_{t\ge0}|\dddot u\ (t)|\le C(\varepsilon,\mathrm{data})$, and thus from $\dot u(t)\to0$ we find $\ddot u(t)\to0$. Therefore $\nabla V(r(t))\to0$ follows from (1.19). □

In the demonstration of the following theorem we use the sublevel sets $\{G_\varepsilon\le c\} = \{(r,u,\dot u) : G_\varepsilon(r,u,\dot u)\le c\}$ and $\{H\le c\} = \{(r,u) : H(r,u)\le c\}$ for $c\in\mathbb{R}$. However, before proceeding, we first have to introduce an appropriate $\bar\delta = \bar\delta(\bar v)>0$ small to modify the functions $a(u)$, $m(u)$, and $b(u,\dot u)$ outside $|u|\le 1-\bar\delta$; cf. section

2. To do this, we assume (U) from now on. The case (U′) is discussed in the remarks below. Since V is bounded and $\bar{v} < 1$, we can find $c_0 \in \mathbb{R}$ such that $\mathcal{K}_{\bar{v}} \subset \{H \leq c_0\}$. Then as a consequence of $\mathcal{E}_s(u) \to \infty$ for $|u| \to 1$, we have

$$s_0 = \sup\big\{|u| : (r,u) \in \{H \leq c_0 + 1\} \text{ for some } r \in \mathbb{R}^3\big\} < 1\,. \tag{4.3}$$

Let us define

$$\bar{\delta} = \min\{(1-\bar{v})/2, (1-s_0)/2\} > 0. \tag{4.4}$$

THEOREM 4.4. *Assume the potential V to satisfy the condition* (U). *Then there exists $\varepsilon_1 > 0$ depending only upon $\bar{v}$ such that for $\varepsilon \in]0,\varepsilon_1]$ all solutions of* (1.19) *starting at points $(r,u,h_\varepsilon(r,u)) \in \mathcal{I}_\varepsilon$, $|u| \leq \bar{v}$, stay away from the boundary $\{(r,u,h_\varepsilon(r,u)) : |u| = 1-\bar{\delta}\}$ for all future times. In particular, solutions exist globally.*

Proof. Let us denote the bound $c(\bar{\delta})$ from (4.1) by c_1 and let us fix $c_a > 0$ such that $|a(u)| \leq c_a$ for all $|u| < 1$. We recall that $a(u)$ was modified to be constant outside $|u| \leq 1-\bar{\delta}$. We define $\varepsilon_1 = \min\{\varepsilon_0, (2c_ac_1)^{-1}\} > 0$.

Let $(r,u) \in \mathcal{K}_{\bar{v}}$. Then $G_\varepsilon(r,u,h_\varepsilon(r,u)) = H(r,u) - \varepsilon(a(u)h_\varepsilon(r,u))\cdot u \leq c_0 + c_ac_1\varepsilon$. Because of Lemma 4.2 the set $\{G_\varepsilon \leq c_0 + c_ac_1\varepsilon\}$ is forward invariant and the solution remains in this set for all future times. On the other hand, since $\bar{v} \leq 1 - 2\bar{\delta} < 1 - \bar{\delta}$, the solution of the modified problem is a solution to (1.19) and stays on $\mathcal{I}_\varepsilon$, at least for a short time. For the fixed time span where this holds the solution is of the form $(r_1,u_1,h_\varepsilon(r_1,u_1))$ and we have $H(r_1,u_1) = G_\varepsilon(r_1,u_1,h_\varepsilon(r_1,u_1)) + \varepsilon(a(u_1)h_\varepsilon(r_1,u_1))\cdot u_1 \leq c_0 + c_ac_1\varepsilon + c_ac_1\varepsilon = c_0 + 2c_ac_1\varepsilon \leq c_0 + 1$ for $\varepsilon \leq \varepsilon_1$. Therefore by (4.3), $|u_1| \leq s_0 \leq 1 - 2\bar{\delta} < 1 - \bar{\delta}$. This argument shows that in fact the solution is confined to $\{(r,u,\dot{u}) : |u| \leq 1 - 2\bar{\delta}\}$. Hence the solution of the modified problem exists, is a solution to (1.19), and stays on $\mathcal{I}_\varepsilon$ for all future times. □

COROLLARY 4.5. *In the setting of Theorem* 4.4, *for solutions of* (1.19) *starting on $\mathcal{I}_\varepsilon$,*

$$\begin{aligned} &\sup\{|u(t)| : t \in \mathbb{R}, \varepsilon \in]0,\varepsilon_1]\} \leq 1 - 2\bar{\delta} < 1\,, \quad \textit{and} \\ &\sup\{|\dot{u}(t)| : t \in \mathbb{R}, \varepsilon \in]0,\varepsilon_1]\} + \sup\{|\ddot{u}(t)| : t \in \mathbb{R}, \varepsilon \in]0,\varepsilon_1]\} \leq c(\bar{\delta}). \end{aligned} \tag{4.5}$$

In particular by Theorem 4.3

$$\dot{u}(t) \to 0, \quad \ddot{u}(t) \to 0, \quad \textit{and} \quad \nabla V(r(t)) \to 0 \quad \textit{as} \quad t \to \infty.$$

Proof. The first estimate was mentioned already in the preceding proof. For the second we note that (4.1) applies, since the trajectory stays on the center manifold, $\dot{u}(t) = h_\varepsilon(r(t),u(t))$. Concerning the last bound, we may write

$$h_\varepsilon(r,u) = -m(u)^{-1}\nabla V(r) + h_{1,\varepsilon}(r,u) \quad \text{with} \quad |h_{1,\varepsilon}(r,u)| \leq c(\bar{\delta})\varepsilon \tag{4.6}$$

for $(r,u) \in \mathbb{R}^3 \times \mathcal{V}$; see [11, Thm. 2.9]. By (1.19),

$$|\varepsilon\ddot{u}| \leq |a(u)^{-1}||m(u)h_{1,\varepsilon}(r,u) - \varepsilon b(u,\dot{u})| \leq c(\bar{\delta})\varepsilon,$$

so we are done. □

Solutions on $\mathcal{I}_\varepsilon$ are uniformly bounded, in the sense of the corollary; in general a bound on $r(t)$ cannot be expected, e.g., in a scattering situation. Conversely, as to

be shown next, solutions with uniformly bounded $u(t)$, $\dot{u}(t)$, and $\ddot{u}(t)$ are confined to the center manifolds.

PROPOSITION 4.6. *Suppose we have a family* $(r^\varepsilon(t), u^\varepsilon(t))$, $\varepsilon \in]0, \varepsilon_2]$, *of solutions to* (1.19) *such that*

$$\sup\{|u^\varepsilon(t)| : t \in \mathbb{R}, \varepsilon \in]0, \varepsilon_2]\} \leq \bar{u} < 1\,, \quad \textit{and}$$
$$\sup\{|\dot{u}^\varepsilon(t)| : t \in \mathbb{R}, \varepsilon \in]0, \varepsilon_2]\} + \sup\{|\ddot{u}^\varepsilon(t)| : t \in \mathbb{R}, \varepsilon \in]0, \varepsilon_2]\} \leq c_2.$$

Then for sufficiently small ε *the solutions have to lie on* $\mathcal{I}_\varepsilon$.

Proof. Note that we can construct $\mathcal{I}_\varepsilon$ here by modifying $a(u), m(u)$, and $b(u, \dot{u})$ to be constant outside, say, $\{u : |u| \leq (1 + \bar{u})/2\}$. According to [11, Thm. 2.1 (ii)] there exists $\delta > 0$ such that for all ε small and solutions $(x(t), y(t))$ to (2.5) the condition $\sup_{t\in\mathbb{R}} |y(t) - h(x(t))| \leq \delta$ implies that the solution has to lie on $\mathcal{I}_\varepsilon$. With $x(t) = (r^\varepsilon(t), u^\varepsilon(t))$ and $y(t) = \dot{u}^\varepsilon(t)$, this condition is verified since we obtain from (1.19) and the assumed bounds $|\dot{u}^\varepsilon(t) + m(u^\varepsilon(t))^{-1}\nabla V(r^\varepsilon(t))| \leq c\varepsilon \leq \delta$, the latter for ε small. □

The asymptotic condition, $\ddot{r}(t) \to 0$, of Dirac [1] and Haag [2] is also sufficient for a solution to lie on $\mathcal{I}_\varepsilon$, in the following sense.

PROPOSITION 4.7. *Suppose a family* $(r^\varepsilon(t), u^\varepsilon(t))$, $\varepsilon \in]0, \varepsilon_2]$, *of solutions to* (1.19) *is given such that*

$$\sup\{|u^\varepsilon(t)| : t \in \mathbb{R}, \varepsilon \in]0, \varepsilon_2]\} \leq \bar{u} < 1$$

and $\ddot{r}^\varepsilon(t) = \dot{u}^\varepsilon(t) \to 0$ *as* $t \to \infty$ *for each* $\varepsilon \in]0, \varepsilon_2]$. *Then for sufficient small* ε *the solutions have to lie on* $\mathcal{I}_\varepsilon$.

Proof. Fix $\delta > 0$. Since Theorem 4.3 applies, we find in the notation of Proposition 4.6

$$|y(t) - h(x(t))| = |\dot{u}^\varepsilon(t) + m(u^\varepsilon(t))^{-1}\nabla V(r^\varepsilon(t))| \leq \delta/2$$

for $t \geq t(\varepsilon)$, with some $t(\varepsilon)$. Thus the solution remains $(\delta/2)$-close to $\mathcal{I}_0$ after time $t(\varepsilon)$, and hence by (4.6) also δ-close to $\mathcal{I}_\varepsilon$ for ε small. Since $\mathcal{I}_\varepsilon$ is normally hyperbolic (repulsive) at an ε-independent rate and since $\delta > 0$ was arbitrary, this can happen only if the solution was already contained in $\mathcal{I}_\varepsilon$. □

Corollary 4.5 provides partial information on the long time behavior of the solutions to (2.5) on the center manifold. Roughly, one can distinguish two classes.

(i) (scattering): The particle enters a domain where $-\nabla V = 0$ at r_1 with velocity u_∞. If the straight line trajectory $r_1 + u_\infty t$, $t \geq 0$ is contained in this domain, then the particle travels freely to infinity. Physically this is a scattering trajectory. In this case $\lim_{t\to\infty} u(t) = u_\infty \neq 0$, whereas the position has no limit.

(ii) (bounded motion): We assume that $|r(t)| \leq$ const. and that within this ball the critical points of V form a discrete set. Then by Corollary 4.5 and by continuity we have $\lim_{t\to\infty} u(t) = 0$ and $\lim_{t\to\infty} r(t) = r_\infty$, where r_∞ is one of the critical points of V. If r_∞ is a stable critical point, then the relaxation is exponentially fast, as can be seen from linearization around the fixed point.

Clearly (i) and (ii) do not exhaust all possibilities. The critical points of V could lie on a sphere. If V is confining, one would still expect convergence to a definite r_∞. Moreover, V could vanish inside a ball. If $-\nabla V$ is pointing towards the ball, then close to each turning point the particle loses energy. Thus $\lim_{t\to\infty} u(t) = 0$, whereas the position has no limit. The potential could decrease so slowly at infinity that no definite velocity is approached. All these cases have to be studied separately.

Up to now we discussed bounded potentials satisfying (U). In the introduction we claimed that our results remain valid also for confining potentials satisfying (U′). In this case, since V is unbounded, we no longer have $\mathcal{K}_{\bar{v}} \subset \{H \leq c_0\}$ for some $c_0 \in \mathbb{R}$ as in Theorem 4.4 above. However, by energy conservation, one can derive the a priori bound $\sup_{t\in\mathbb{R}} |q^\varepsilon(t)| \leq \bar{M}$ for solutions to the true system (1.5) on the macroscopic time scale. Thus the motion is bounded also in the q-direction and it suffices to build the center manifold for the effective equation (1.19) over the bounded domain

$$\mathcal{K}_{\bar{M},\bar{v}} = \{r \in \mathbb{R}^3 : |r| \leq \bar{M}\} \times \{u \in \mathbb{R}^3 : |u| \leq \bar{v}\},$$

enlarged to a suitable $\mathcal{K}_{\bar{M}+1,1-\bar{\delta}}$ such that solutions starting over $\mathcal{K}_{\bar{M},\bar{v}}$ stay away from the boundary of $\mathcal{K}_{\bar{M}+1,1-\bar{\delta}}$ for $\varepsilon > 0$ sufficiently small. In this manner we obtain

THEOREM 4.8. *Assume* (U′) *holds for the potential, and let* $\mathcal{K}_{\bar{M},\bar{\delta}}$ *be defined as above. Then there exists* $\varepsilon_1 > 0$ *depending only on the initial data such that for* $\varepsilon \in]0, \varepsilon_1]$ *all solutions of* (1.19) *starting at points* $(r, u, h_\varepsilon(r, u)) \in \mathcal{I}_\varepsilon$, $(r, u) \in \mathcal{K}_{\bar{M},\bar{\delta}}$, *exist globally. Moreover, these solutions are uniformly bounded:*

$$\begin{aligned} \sup\{|r(t)| : t \in \mathbb{R}, \varepsilon \in]0, \varepsilon_1]\} &\leq c(\bar{\delta}), \\ \sup\{|u(t)| : t \in \mathbb{R}, \varepsilon \in]0, \varepsilon_1]\} &\leq 1 - 2\bar{\delta} < 1, \quad \textit{and} \\ \sup\{|\dot{u}(t)| : t \in \mathbb{R}, \varepsilon \in]0, \varepsilon_1]\} + \sup\{|\ddot{u}(t)| : t \in \mathbb{R}, \varepsilon \in]0, \varepsilon_1]\} &\leq c(\bar{\delta}). \end{aligned} \tag{4.7}$$

In addition,

$$\dot{u}(t) \to 0, \quad \ddot{u}(t) \to 0, \quad \textit{and} \quad \nabla V(r(t)) \to 0 \quad \textit{as} \quad t \to \infty.$$

Proof. The proof is similar to the one of Theorem 4.4. Concerning the boundedness, note that again for some $c_0 \in \mathbb{R}$ and $\varepsilon > 0$ small,

$$\mathcal{K}_{\bar{M},\bar{\delta}} \subset \{H \leq c_0\} \subset \{(r, u) : G_\varepsilon(r, u, h_\varepsilon(r, u)) \leq c_0 + c_a c_1 \varepsilon\} \subset \{H \leq c_0 + 1\}.$$

Thus all solutions starting over $\mathcal{K}_{\bar{M},\bar{\delta}}$ will remain on the manifolds over $\{H \leq c_0 + 1\}$. Since this set is independent of ε and compact by (U′), the solutions must be uniformly bounded, because h_ε is uniformly bounded. □

On the center manifold the motion is governed by the (second order) equation

$$\dot{x} = f(x, h_\varepsilon(x)). \tag{4.8}$$

Since the existence of h_ε is established only abstractly, (4.8) is somewhat implicit. From [11, (2.9-1) and Thm. 2.9] we know that h_ε depends smoothly on ε. Thus (4.8) can be expanded in ε. Including the first Taylor term we pick up an error of order ε^2, which is of the same order as the error between the true and the comparison dynamics on the center manifold. For consistency we should stop then at this order. We make the ansatz

$$h_\varepsilon(r, u) = h_0(r, u) + \varepsilon h_1(r, u) + h_{2,\varepsilon}(r, u), \quad |h_{2,\varepsilon}(r, u)| \leq c(\bar{\delta})\varepsilon^2.$$

for $(r, u) \in \mathbb{R}^3 \times \mathcal{V}$. Then

$$m(u)h_0(r, u) = -\nabla V(r),$$

and $h_1(r, u)$ is determined through

$$D_x h_0(r, u) f(r, u, h_0(r, u)) = D_y g(r, u, h_0(r, u), 0) h_1(r, u) + D_\varepsilon g(r, u, h(r, u), 0);$$

see [11, (2.9-1) and Thm. 2.9]. Computing the respective derivatives one arrives at

$$\begin{aligned} m(u)h_1(r,u) = a(u)\Big[&- m(u)^{-1}\nabla^2 V(r)u \\ &+\Big(\frac{d}{du}m(u)\Big)^{-1}\big(\nabla V(r), m(u)^{-1}\nabla V(r)\big) \\ &+ b(u, m(u)^{-1}\nabla V(r))\Big] \end{aligned}$$

and the effective second order equation

$$\dot r = u\,, \quad m(u)\dot u = -\nabla V(r) + \varepsilon m(u)h_1(r,u) \tag{4.9}$$

of the particle motion on the center manifold.

5. Comparison of the true and the effective system. In this section we prove Theorem 2.2. Since we have $|u(t_1)| = |v^\varepsilon(t_1)| \le \bar v$ by (1.8), Theorem 4.4, resp., Theorem 4.8, implies that the solution trajectory of the system with the modified functions $a(u)$, $m(u)$, and $b(u,\dot u)$ is indeed a solution trajectory to (1.19). Recall that, by (1.18) and (1.19),

$$m(v^\varepsilon)\dot v^\varepsilon = -\nabla V(q^\varepsilon) + \varepsilon a(v^\varepsilon)\ddot v^\varepsilon + \varepsilon b(v^\varepsilon,\dot v^\varepsilon) + \varepsilon^2 f^\varepsilon(t), \quad t \ge \varepsilon t_1, \tag{5.1}$$

$$m(u)\dot u = -\nabla V(r) + \varepsilon a(u)\ddot u + \varepsilon b(u,\dot u), \tag{5.2}$$

with $|f^\varepsilon(t)| \le C$. Using the bounds (1.8) and (4.5), resp., (4.7), we infer the weaker estimate

$$\begin{aligned} m(v^\varepsilon)\dot v^\varepsilon &= -\nabla V(q^\varepsilon) + \mathcal{O}(\varepsilon), \quad t \ge t_1, \\ m(u)\dot u &= -\nabla V(r) + \mathcal{O}(\varepsilon), \end{aligned}$$

which has been proved already in [5]. Hence (2.7) follows by the argument therein.

To show (2.8) we compute as in Lemma 4.2, using (5.1),

$$\begin{aligned} &\frac{d}{dt}G_\varepsilon(q^\varepsilon(t), v^\varepsilon(t), \dot v^\varepsilon(t)) \\ &= \varepsilon^2 f^\varepsilon(t)v^\varepsilon(t) \\ &\quad -\varepsilon(\mathsf{e}^2/12\pi)\Big[\gamma(\dot v^\varepsilon(t))^6\dot v^\varepsilon(t)^2 + 6\gamma(\dot v^\varepsilon(t))^8(v^\varepsilon(t)\cdot\dot v^\varepsilon(t))^2\Big], \quad t \ge \varepsilon t_1. \end{aligned}$$

Since $r(\varepsilon t_1) = q^\varepsilon(\varepsilon t_1)$ and $u(\varepsilon t_1) = v^\varepsilon(\varepsilon t_1)$, using the uniform bounds, we have for $t \ge \varepsilon t_1$

$$\begin{aligned} &|H(q^\varepsilon(t), v^\varepsilon(t)) - H(r(t),u(t))| \\ &\le \varepsilon\big|(a(v^\varepsilon(t))\dot v^\varepsilon(t))\cdot v^\varepsilon(t) - (a(u(t))\dot u(t))\cdot u(t)\big| \\ &\quad + \int_{\varepsilon t_1}^t ds\,\Big[\varepsilon^2|f^\varepsilon(s)v^\varepsilon(s)| + \varepsilon(\mathsf{e}^2/12\pi)\big(\big|\gamma(\dot v^\varepsilon(s))^6\dot v^\varepsilon(s)^2 - \gamma(\dot u(s))^6\dot u^\varepsilon(s)^2\big| \\ &\qquad\qquad +6\,\big|\gamma(\dot v^\varepsilon(s))^8(v^\varepsilon(s)\cdot\dot v^\varepsilon(s))^2 - \gamma(\dot u(s))^8(u(s)\cdot\dot u(s))^2\big|\big)\Big] \\ &\le C\varepsilon\big[|v^\varepsilon(t) - u(t)| + |\dot v^\varepsilon(t) - \dot u(t)|\big] + C\varepsilon^2 t \\ &\quad + C\varepsilon\int_{\varepsilon t_1}^t ds\,\big[|v^\varepsilon(s) - u(s)| + |\dot v^\varepsilon(s) - \dot u(s)|\big] \\ &\le C\varepsilon^2(1+t) \le C\varepsilon^2, \end{aligned}$$

by (2.7) for $t = \mathcal{O}(1)$. This concludes the proof of Theorem 2.2. ∎

Finally we show that on a microscopic time scale our results track the true trajectory with a higher precision; cf. (2.9), Remark 2.3(i).

PROPOSITION 5.1. *We have*

$$|q^\varepsilon(t) - r(t)| \le c\varepsilon^3 \quad \textit{and} \quad |v^\varepsilon(t) - u(t)| \le c\varepsilon^2, \quad t = \mathcal{O}(\varepsilon),$$

i.e., (2.9) *holds.*

Proof. Define $\Psi(s) = \big(\varepsilon^{-1}q^\varepsilon(\varepsilon s) - \varepsilon^{-1}r(\varepsilon s), v^\varepsilon(\varepsilon s) - u(\varepsilon s), \varepsilon\dot v^\varepsilon(\varepsilon s) - \varepsilon\dot u(\varepsilon s)\big)$ for $s \ge t_1$. Then $\dot\Psi(s) = A\Psi(s) + \theta(s)$, where

$$A = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

and $\theta(s) = \varepsilon^2(0, 0, \ddot v^\varepsilon(\varepsilon s) - \ddot u(\varepsilon s))$, from which

$$\begin{aligned} |\theta(s)| &\le c\varepsilon \, \big(|q^\varepsilon(\varepsilon s) - r(\varepsilon s)| + |v^\varepsilon(\varepsilon s) - u(\varepsilon s)| + |\dot v^\varepsilon(\varepsilon s) - \dot u(\varepsilon s)| + \varepsilon^2\big) \\ &\le c(|\Psi(s)| + \varepsilon^3), \quad s \ge t_1, \end{aligned}$$

by (5.1), (5.2), and the uniform bounds. Therefore by the variation of constants formula and Gronwall's inequality for $s \in [t_1, t_1 + \tau]$, $|\Psi(s)| \le c(\tau)(|\Psi(s_1)| + \varepsilon^3)$. Consequently, $q^\varepsilon(\varepsilon t_1) = r(\varepsilon t_1)$ and $v^\varepsilon(\varepsilon t_1) = u(\varepsilon t_1)$ yields

$$\varepsilon^{-1}|q^\varepsilon(t) - r(t)| + |v^\varepsilon(t) - u(t)| \le c(\tau)\, \big(\varepsilon|\dot v^\varepsilon(\varepsilon t_1) - \dot u(\varepsilon t_1)| + \varepsilon^3\big)$$

for $t \in [\varepsilon t_1, \varepsilon t_1 + \varepsilon\tau]$. By (5.1) and (5.2), $|\dot v^\varepsilon(\varepsilon t_1) - \dot u(\varepsilon t_1)| \le c\varepsilon$, and therefore (2.9) follows. ∎

6. Adiabatic limit of the fields. We prove Theorem 2.4. Let $U(t)$ again denote the fundamental solution of the wave equation in $D^{1,2}(\mathbb{R}^3) \oplus L^2(\mathbb{R}^3)$. We set $Z(x,t) = (\phi(x,t), \pi(x,t))$ as well as $\bar\rho_\varepsilon = (0, \rho_\varepsilon)$. Then the inhomogeneous wave equation in (1.5) is solved as

$$Z(x,t) = [U(t)Z(\cdot,0)](x) - \sqrt{\varepsilon} \int_0^t ds\, [U(t-s)\bar\rho_\varepsilon(\cdot - q^\varepsilon(s))](x).$$

Since

$$Z(x,0) = -\sqrt{\varepsilon} \int_{-\infty}^0 ds\, [U(-s)\bar\rho_\varepsilon(\cdot - q^0 - v^0 s)](x)$$

(cf. Lemma 3.1), we have for $t > 0$

$$Z(x,t) = -\sqrt{\varepsilon} \int_{-\infty}^t ds\, [U(t-s)\bar\rho_\varepsilon(\cdot - q^\varepsilon(s))](x),$$

where we extended the position to negative times $t \le 0$ by $q^\varepsilon(t) = q^0 + v^0 t$. Thus by the solution formula for the wave equation

$$\frac{1}{\sqrt{\varepsilon}}\, \phi(x,t) = -\int \frac{d^3y}{4\pi|x-y|}\, \rho_\varepsilon(y - q^\varepsilon(t - |x-y|)) \tag{6.1}$$

and $\pi(x,t) = \dot\phi(x,t)$. For $\varepsilon \to 0$, $q^\varepsilon(t) \to r(t)$ (cf. (1.9)), with $r(t)$ extended to negative times by $r(t) = q^0 + v^0 t$. Moreover, $\rho_\varepsilon(x) = \varepsilon^{-3}\rho(\varepsilon^{-1}x) \to \mathsf{e}\delta_0$ in the sense of distributions. Hence the transformation $z = y - q^\varepsilon(t - |x-y|)$, $\det(dy/dz) = [1 - v^\varepsilon(t-|x-y|)\cdot(x-y)/|x-y|]^{-1}$, in (6.1) yields the pointwise convergence (2.10), except on the worldline of the particle, since the integrand in (6.1) is singular at $y = x$, i.e., for $x = r(t)$ which corresponds to $t_{\rm ret} = t$.

The analogous argument works for $\pi(x,t)$. In the limit $\varepsilon \to 0$, π is discontinuous at the light cone $\{x : |x| = t\}$, which we avoided due to our assumption. □

7. Radiated energy. Let $E_{R,q^\varepsilon(t)}(t+R)$ be the energy, particle plus field, at time $t+R$ in a ball of radius R centered at $q^\varepsilon(t)$. For $R > \varepsilon R_\rho$ this energy changes as

$$\begin{aligned}
&\frac{d}{dt}\,\big(E_{R,q^\varepsilon(t)}(t+R)\big)\\
&= \frac{d}{dt}\bigg(\mathcal{H}_{mac}(t=0) - \frac{1}{2}\int_{\{|x-q^\varepsilon(t)|>R\}} d^3x\,\big[|\pi(x,t+R)|^2 + |\nabla\phi(x,t+R)|^2\big]\bigg)\\
&= R^2\int_{|\omega|=1} d^2\omega\,\pi(q^\varepsilon(t)+R\omega,t+R)\,\omega\cdot\nabla\phi(q^\varepsilon(t)+R\omega,t+R)\\
&\quad + \frac{R^2}{2}\int_{|\omega|=1} d\omega\,(\omega\cdot v^\varepsilon(t))\big[|\pi(q^\varepsilon(t)+R\omega,t+R)|^2\\
&\qquad\qquad + |\nabla\phi(q^\varepsilon(t)+R\omega,t+R)|^2\big], \qquad (7.1)
\end{aligned}$$

where we used that the total energy is conserved.

E_R changes because there is energy flowing back and forth between particle and field, and because energy is lost irreversibly to infinity. To separate both contributions we take the limit $R \to \infty$. Using (6.1) and the relation $t + R - |q^\varepsilon(t) + R\omega - y| = t + \omega\cdot(y - q^\varepsilon(t)) + \mathcal{O}(1/R)$ for bounded $|y|$, we arrive at

$$\begin{aligned}
I^\varepsilon(t) &= \lim_{R\to\infty}\frac{d}{dt}\,\big(E_{R,q^\varepsilon(t)}(t+R)\big)\\
&= -\varepsilon(4\pi)^{-2}\int_{|\omega|=1} d^2\omega\,(1-\omega\cdot v^\varepsilon(t))\,\bigg[\int d^3y\,\rho_\varepsilon(y - q^\varepsilon(t+\omega\cdot[y-q^\varepsilon(t)]))\\
&\qquad\qquad \times\frac{\omega\cdot\dot v^\varepsilon(t+\omega\cdot[y-q^\varepsilon(t)])}{(1-\omega\cdot v^\varepsilon(t+\omega\cdot[y-q^\varepsilon(t)]))^2}\bigg]^2;
\end{aligned}$$

cf. [7, sec. 3] for details on a similar calculation. In fact, in [7] the ball of radius R was centered at the origin and the second summand in (7.1) is absent. To let $\varepsilon \to 0$, we again transform to $z = y - q^\varepsilon(t+\omega\cdot[y-q^\varepsilon(t)])$, $\det(dy/dz) = [1 - \omega\cdot v^\varepsilon(t+\omega\cdot[y-q^\varepsilon(t)])]^{-1}$, use $\rho_\varepsilon(x) \to \mathsf{e}\delta_0$ in the sense of distributions, and insert the identity $y = q^\varepsilon(t)$ for $z = 0$ to obtain

$$\begin{aligned}
\lim_{\varepsilon\to 0}\varepsilon^{-1}\,I^\varepsilon(t) &= -\mathsf{e}^2(4\pi)^{-2}\int_{|\omega|=1} d^2\omega\,(1-\omega\cdot u(t))^{-5}\,(\omega\cdot\dot u(t))^2\\
&= -(\mathsf{e}^2/12\pi)\,[6\gamma^8(u(t)\cdot\dot u(t))^2 + \gamma^6\dot u(t)^2],
\end{aligned}$$

in agreement with (4.2).

Alternatively, we could first take the limit $\varepsilon \to 0$ in (7.1). Using Theorem 2.4 we find, with $(\overline\phi, \overline\pi)$ denoting the limit fields from (2.10), (2.11),

$$
\begin{aligned}
I_R(t) &= \lim_{\varepsilon\to 0} \varepsilon^{-1} \frac{d}{dt}\left(E_{R,q^\varepsilon(t)}(t+R)\right) \\
&= R^2 \int_{|\omega|=1} d^2\omega\, \overline{\pi}(r(t)+R\omega, t+R)\, \omega\cdot\nabla\overline{\phi}(r(t)+R\omega, t+R) \\
&\quad + \frac{R^2}{2}\int_{|\omega|=1} d^2\omega\, (\omega\cdot u(t))\left[|\overline{\pi}(r(t)+R\omega,t+R)|^2\right. \\
&\qquad\qquad \left. +|\nabla\overline{\phi}(r(t)+R\omega,t+R)|^2\right].
\end{aligned}
$$

Since both $\overline{\pi}$ and $\nabla\overline{\phi}$ have one term proportional to R^{-1} and other contributions of order R^{-2}, in the limit $R\to\infty$ only the product of the two leading terms survives, and it follows that

$$
\lim_{R\to\infty} I_R(t) = -\mathsf{e}^2(4\pi)^{-2}\int_{|\omega|=1} d^2\omega\, (1-\omega\cdot u(t))^{-5}\,(\omega\cdot\dot{u}(t))^2,
$$

as before. We note that the radiated energy is of order ε and it therefore suffices to use the effective dynamics to order one, i.e., ignoring the radiation reaction.

8. Appendix: Proof of Lemma 2.1. In this appendix we prove Lemma 2.1. Since we need to use some identities from [5], we switch back to the original time scale of (1.1). Hence we have to show the following.

LEMMA 2.1 *For solutions of* (1.1) *with initial values satisfying* (1.4)*, i.e., starting on the soliton manifold, and for* $|\rho|_{L^2}$ *sufficiently small we have*

$$
\sup_{t\in\mathbb{R}} |\ddot{v}\ (t)| \le C\varepsilon^3\,.
$$

The constant C *and the bound on* $|\rho|_{L^2}$ *depend only on the data.*

Proof. From [5, Lem. 2.2 and Prop. 4.1] we already know the bounds

$$
\begin{aligned}
&\sup_{t\in\mathbb{R}}|v(t)| \le \bar{v} < 1\,, \quad \sup_{t\in\mathbb{R}}|\dot{v}(t)| + \sup_{t\in\mathbb{R}}|\dot{p}(t)| \le C\varepsilon, \\
&\qquad\text{and}\quad \sup_{t\in\mathbb{R}}|\ddot{v}(t)| + \sup_{t\in\mathbb{R}}|\ddot{p}(t)| \le C\varepsilon^2
\end{aligned}
\tag{8.1}
$$

for $|\rho|_{L^2}$ sufficiently small. The constants $\bar{v}$ and C appearing in (8.1) do not depend on the particular solution, but only on bounds for the initial values.

Denote

$$
Z(x,t) = \begin{pmatrix} \varphi(x,t) \\ \psi(x,t)\end{pmatrix} = \begin{pmatrix} \phi(x,t) - \phi_{v(t)}(x-q(t)) \\ \pi(x,t) - \pi_{v(t)}(x-q(t))\end{pmatrix}.
$$

Then (cf. [5]), with $L(t)\phi = \nabla\phi\cdot v(t) + \dot{\phi}$,

$$
\begin{aligned}
\ddot{p}(t) &= -\varepsilon^2\nabla^2 V(\varepsilon q(t))\cdot v(t) + \int d^3x\,(L(t)\varphi)(x+q(t),t)\nabla\rho(x) \\
&=: -\varepsilon^2\nabla^2 V(\varepsilon q(t))\cdot v(t) + M(t).
\end{aligned}
$$

Therefore

$$
\dddot{p}\ (t) = -\varepsilon^2\nabla^2 V(\varepsilon q(t))\cdot\dot{v}(t) - \varepsilon^3\nabla^3 V(\varepsilon q(t))(v(t),v(t)) + \dot{M}(t)\,. \tag{8.2}
$$

Below we will show the following.

LEMMA 8.1. *The estimate*

$$|\dot{M}(t)| \le C\left(\varepsilon^3 + |\rho|_{L^2}\int_0^t \frac{|\ \dddot{v}\ (s)|}{1+(t-s)^2}\,ds\right)$$

holds.

Then according to (8.2), (8.1), and assumption (U) on the potential,

$$|\ \ddot{p}\ (t)| \le C\left(\varepsilon^3 + |\rho|_{L^2}\int_0^t \frac{|\ \dddot{v}\ (s)|}{1+(t-s)^2}\,ds\right). \tag{8.3}$$

Since

$$\begin{aligned}|\ \ddot{v}\ | &= \left|\frac{d}{dp}\left(\frac{p}{\sqrt{1+p^2}}\right)\ddot{p} + 3\,\frac{d^2}{dp^2}\left(\frac{p}{\sqrt{1+p^2}}\right)(\dot{p},\ddot{p}) + \frac{d^3}{dp^3}\left(\frac{p}{\sqrt{1+p^2}}\right)(\dot{p},\dot{p},\dot{p})\right| \\ &\le C(|\ \dddot{p}\ | + \varepsilon^3),\end{aligned}$$

the claim of Lemma 2.1 is obtained from (8.3) by taking $|\rho|_{L^2}$ small enough. □

Thus it remains to give the following.

Proof of Lemma 8.1. First note

$$\dot{M}(t) = \int d^3x\,\langle(\mathcal{L}(t)Z(\cdot,t))(x),\,\nabla\rho_*(x-q(t))\rangle_{\mathbb{R}^2}\,,\quad \rho_*(x) = (\rho(x),0),$$

where $\mathcal{L}(t)Z = \nabla Z\cdot\dot{v}(t) + (\nabla^2 Z)(v(t),v(t)) + 2\nabla\dot{Z}\cdot v(t) + \ddot{Z}$. Because $\dot{Z} = AZ - B$, with

$$A(\phi,\pi) = (\pi,\Delta\phi)\quad\text{and}\quad B(x,t) = \begin{pmatrix}\nabla_v\phi_{v(t)}(x-q(t))\cdot\dot{v}(t)\\ \nabla_v\pi_{v(t)}(x-q(t))\cdot\dot{v}(t)\end{pmatrix}, \tag{8.4}$$

we obtain

$$\frac{d}{dt}(\mathcal{L}(t)Z) = A(\mathcal{L}(t)Z) - \mathcal{L}(t)B + 2\left[(\nabla^2 Z)(v,\dot{v}) + \nabla\dot{Z}\cdot\dot{v}\right] + \nabla Z\cdot\ddot{v}\,.$$

Let $U(t)$ again denote the group generated by the free wave equation on $D^{1,2}(\mathbb{R}^3)\oplus L^2(\mathbb{R}^3)$. Then

$$\begin{aligned}\dot{M}(t) \;=\;& \langle U(t)[\mathcal{L}(0)Z(\cdot,0)],\,\nabla\rho_*(\cdot-q(t))\rangle_{L^2(\mathbb{R}^2)} \\ &+\int_0^t ds\,\Big[\,-\langle U(t-s)[\mathcal{L}(s)B(\cdot,s)],\,\nabla\rho_*(\cdot-q(t))\rangle_{L^2(\mathbb{R}^2)} \\ &\qquad+2\,\langle U(t-s)[(\nabla^2 Z(\cdot,s))(v(s),\dot{v}(s)) + \nabla\dot{Z}(\cdot,s)\cdot\dot{v}(s)], \\ &\qquad\qquad\nabla\rho_*(\cdot-q(t))\rangle_{L^2(\mathbb{R}^2)} \\ &\qquad+\langle U(t-s)[\nabla Z(\cdot,s)\cdot\ddot{v}(s)],\,\nabla\rho_*(\cdot-q(t))\rangle_{L^2(\mathbb{R}^2)}\,\Big] \\ =:\;& T_0+T_1+T_2+T_3\,.\end{aligned}$$

We estimate each term T_j separately, keeping all parts which contain only initial values. Note that here according to (1.4) we have $Z(x,0)=0$, so all these terms vanish. Nevertheless, we wanted to derive the general form of the estimate; see Remark 2.3(iii).

Estimate of T_3: Since $\dot{Z} = AZ - B$, we find

$$Z(t) = U(t)Z(0) - \int_0^t ds\,U(t-s)B(\cdot,s), \tag{8.5}$$

and hence

$$\begin{aligned} T_3 &= \int_0^t ds\, \langle U(t)[\nabla Z(\cdot,0)\cdot \ddot{v}(s)],\, \nabla\rho_*(\cdot - q(t))\rangle_{L^2(\mathbb{R}^2)} \\ &\quad - \int_0^t ds \int_0^s d\tau\, \langle U(t-\tau)[\nabla B(\cdot,\tau)\cdot \ddot{v}(s)],\, \nabla\rho_*(\cdot - q(t))\rangle_{L^2(\mathbb{R}^2)} \\ &=: T_{3,0} + T_{3,1}\,. \end{aligned}$$

Then Lemma 8.2 below and (8.1) imply through integration by parts in the d^3x-integral

$$T_{3,1} \le C\varepsilon^2 \int_0^t ds \int_0^s d\tau\, \frac{\varepsilon}{1+(t-\tau)^3} \le C\varepsilon^3\,.$$

Estimate of T_0: This term is determined solely through the data.

Estimate of T_1: If we calculate the form of $\mathcal{L}(t)B = \nabla B\cdot \dot{v} + (\nabla^2 B)(v,v) + 2\nabla\dot{B}\cdot v + \ddot{B}$ explicitly from (8.4), many terms cancel, fortunately, and we find with $\Phi_v = \binom{\phi_v}{\pi_v}$,

$$\mathcal{L}(t)B = \nabla_v\Phi_v(x-q)\cdot\, \dddot{v} + 3\nabla_v^2\Phi_v(x-q)(\dot{v},\ddot{v}) + \nabla_v^3\Phi_v(x-q)(\dot{v},\dot{v},\dot{v})\,.$$

Now we may argue analogously to the estimate of T_3 and Lemma 8.2 to obtain with

$$\begin{pmatrix} \tilde{\phi}(x) \\ \tilde{\pi}(x) \end{pmatrix} = [U(t-s)\mathcal{L}(s)B(\cdot,s)](x)$$

the estimate

$$|\nabla\tilde{\phi}(x+q(t))| \le \frac{C}{1+(t-s)^2}\left(\varepsilon^3 + |\,\dddot{v}\,(s)|\right), \quad |x|\le R_\rho\,, \quad t\ge s\,. \tag{8.6}$$

Here we have used (8.1) and some of the estimates

$$\begin{aligned} |\nabla\nabla_v\phi_v(x)| + |\nabla\nabla_v^2\phi_v(x)| + |\nabla\nabla_v^3\phi_v(x)| &\le C(1+|x|)^{-2}\,, \\ |\nabla^2\nabla_v\phi_v(x)| + |\nabla^2\nabla_v^2\phi_v(x)| + |\nabla^2\nabla_v^3\phi_v(x)| &\le C(1+|x|)^{-3}\,, \\ |\nabla^3\nabla_v\phi_v(x)| + |\nabla^3\nabla_v^2\phi_v(x)| + |\nabla^3\nabla_v^3\phi_v(x)| &\le C(1+|x|)^{-4}\,, \\ |\nabla^4\nabla_v\phi_v(x)| + |\nabla^4\nabla_v^2\phi_v(x)| + |\nabla^4\nabla_v^3\phi_v(x)| &\le C(1+|x|)^{-5}\,, \\ |\nabla\nabla_v\pi_v(x)| + |\nabla\nabla_v^2\pi_v(x)| + |\nabla\nabla_v^3\pi_v(x)| &\le C(1+|x|)^{-3}\,, \\ |\nabla^2\nabla_v\pi_v(x)| + |\nabla^2\nabla_v^2\pi_v(x)| + |\nabla^2\nabla_v^3\pi_v(x)| &\le C(1+|x|)^{-4}\,, \\ |\nabla^3\nabla_v\pi_v(x)| + |\nabla^3\nabla_v^2\pi_v(x)| + |\nabla^3\nabla_v^3\pi_v(x)| &\le C(1+|x|)^{-5} \end{aligned} \tag{8.7}$$

for $x\in\mathbb{R}^3$ and $|v|\le\bar{v}$. From (8.6) we conclude

$$T_1 = -\int_0^t ds \int_{|x|\le R_\rho} d^3x\, \nabla\tilde{\phi}(x+q(t))\rho(x) \le C\left(\varepsilon^3 + |\rho|_{L^2}\int_0^t \frac{|\,\dddot{v}\,(s)|}{1+(t-s)^2}\,ds\right).$$

Estimate of T_2: Let $P(t)Z = \nabla^2 Z(\cdot,t)v(t) + \nabla\dot{Z}(\cdot,t)$. Then

$$\frac{d}{dt}(P(t)Z) = P(t)\dot{Z} + (\nabla^2 Z)\dot{v} = A(P(t)Z) - P(t)B + (\nabla^2 Z)\dot{v}\,.$$

Therefore by the definition of T_2,

$$\begin{aligned} T_2 \;=\; & 2\int_0^t ds\,\langle U(t)[(P(0)Z(\cdot,0))\cdot\dot v(s)],\,\nabla\rho_*(\cdot-q(t))\rangle_{L^2(\mathbb{R}^2)} \\ & +2\int_0^t ds\int_0^s d\tau\,\langle U(t-\tau)\left[-P(\tau)B(\cdot,\tau)+(\nabla^2 Z(\cdot,\tau))\dot v(\tau)\right]\cdot\dot v(s), \\ & \qquad\qquad \nabla\rho_*(\cdot-q(t))\rangle_{L^2(\mathbb{R}^2)} \\ =:\; & T_{2,0}+T_{2,1}+T_{2,2}\,. \end{aligned}$$

To estimate $T_{2,1}$, observe

$$P(t)B=\nabla\nabla_v\Phi_v(x-q)\cdot\ddot v+\nabla\nabla_v^2\Phi_v(x-q)(\dot v,\dot v)\,.$$

Hence we may argue as before to find $|T_{2,1}|\le C\varepsilon^3$. In order to bound $T_{2,2}$, similarly to the estimate of T_3, we again use (8.5) to get

$$\begin{aligned} & T_{2,2} \\ & \quad = 2\int_0^t ds\int_0^s d\tau\,\langle U(t)[\nabla^2 Z(0)(\dot v(\tau),\dot v(s))],\,\nabla\rho_*(\cdot-q(t))\rangle_{L^2(\mathbb{R}^2)} \\ & \qquad -2\int_0^t ds\int_0^s d\tau\int_0^\tau d\sigma\,\langle U(t-\sigma)[\nabla^2 B(\cdot,\sigma)(\dot v(\tau),\dot v(s))], \\ & \qquad\qquad\qquad \nabla\rho_*(\cdot-q(t))\rangle_{L^2(\mathbb{R}^2)} \\ & \quad =: T_{2,2,0}+T_{2,2,1}\,. \end{aligned}$$

By (8.7) and the argument of Lemma 8.2 then

$$T_{2,2,1}\le\int_0^t ds\int_0^s d\tau\int_0^\tau d\sigma\,\frac{C\varepsilon^3}{1+(t-\sigma)^4}\le C\varepsilon^3\,.$$

Summarizing all above estimates for T_0–T_3, we hence arrive at

$$\begin{aligned} |\dot M(t)|\le\; & C\left(\varepsilon^3+|\rho|_{L^2}\int_0^t\frac{|\;\ddot v\;(s)|}{1+(t-s)^2}\,ds\right) \\ & +\langle U(t)[\mathcal{L}(0)Z(\cdot,0)],\,\nabla\rho_*(\cdot-q(t))\rangle_{L^2(\mathbb{R}^2)} \\ & +2\int_0^t ds\,\langle U(t)[(P(0)Z(\cdot,0))\cdot\dot v(s)],\,\nabla\rho_*(\cdot-q(t))\rangle_{L^2(\mathbb{R}^2)} \\ & +2\int_0^t ds\int_0^s d\tau\,\langle U(t)[\nabla^2 Z(0)(\dot v(\tau),\dot v(s))],\,\nabla\rho_*(\cdot-q(t))\rangle_{L^2(\mathbb{R}^2)} \\ & +\int_0^t ds\,\langle U(t)[\nabla Z(\cdot,0)\cdot\ddot v(s)],\,\nabla\rho_*(\cdot-q(t))\rangle_{L^2(\mathbb{R}^2)}\,. \end{aligned} \tag{8.8}$$

Concerning the terms that contain data, these vanish here since $Z(x,0)=0$ as a consequence of (1.4). This completes the proof of Lemma 8.1. □

In case of solutions starting close to but not on the soliton manifold as discussed in Remark 2.3(iii), conditions on the data have to be imposed to ensure the last four terms in (8.8) can also be estimated by $C\varepsilon^3$. In [5, Thm. 2.6 and sect. 4] details are carried out for derivatives of one order less.

We used the following lemma above.

LEMMA 8.2. *The estimate*

$$\|\nabla[U(t-\tau)\nabla B(\cdot,\tau)](\cdot+q(t))\|_{R_\rho} \leq C\,\frac{\varepsilon}{1+(t-\tau)^3}\,, \quad t\geq\tau\,, \tag{8.9}$$

holds.

Proof. Such estimates have already been used in [5], but we nevertheless include some details of the argument. Let

$$\begin{pmatrix} \tilde\phi(x) \\ \tilde\pi(x) \end{pmatrix} = [U(t-\tau)\nabla B(\cdot,\tau)](x)$$

for fixed t,τ. By Kirchhoff's formula for the solution to the wave equation and by (8.4),

$$\begin{aligned} &\nabla\tilde\phi(x+q(t)) \\ &= \frac{1}{4\pi(t-\tau)^2}\int_{|y-x-q(t)|=(t-\tau)} d^2y\,\big[(t-\tau)\nabla^2\nabla_v\pi_{v(\tau)}(y-q(\tau))\cdot\dot v(\tau) \\ &\qquad +\nabla^2\nabla_v\phi_{v(\tau)}(y-q(\tau))\cdot\dot v(\tau) \\ &\qquad +\nabla^3\nabla_v\phi_{v(\tau)}(y-q(\tau))(\dot v(\tau),y-x-q(t))\,\big]. \end{aligned} \tag{8.10}$$

Now $|x|\leq R_\rho$ and $|y-x-q(t)|=(t-\tau)$ yield $|y-q(\tau)|\geq(t-\tau)-\bar v(t-\tau)-R_\rho = (1-\bar v)(t-\tau)-R_\rho$ by (8.1). As a consequence of (8.7), (8.9) therefore follows from (8.10). □

Acknowledgment. We thank A. Komech for useful discussions.

REFERENCES

[1] P. A. M. DIRAC, *Classical theory of radiating electrons*, Proc. Royal Soc. London Ser. A, 167 (1938), pp. 148–169.
[2] R. HAAG, *Die Selbstwechselwirkung des Elektrons*, Z. Naturforsch., 10a (1955), pp. 752–761.
[3] J. D. JACKSON, *Classical Electrodynamics*, 2nd ed., John Wiley and Sons, New York, London, 1975.
[4] CH. JONES, *Geometric singular perturbation theory*, in Dynamical Systems, Proceedings, Montecatini Terme 1994, Lecture Notes in Math. 1609, R. Johnson, ed., Springer-Verlag, Berlin, New York, 1995, pp. 44–118.
[5] A. KOMECH, M. KUNZE, AND H. SPOHN, *Effective dynamics for a mechanical particle coupled to a wave field*, Comm. Math. Phys., 203 (1999), pp. 1–19.
[6] A. KOMECH AND H. SPOHN, *Soliton-like asymptotics for a classical particle interacting with a scalar wave field*, Nonlinear Anal., 33 (1998), pp. 13–24.
[7] A. KOMECH, H. SPOHN, AND M. KUNZE, *Long-time asymptotics for a classical particle interacting with a scalar wave field*, Comm. Partial Differential Equations, 22 (1997), pp. 307–335.
[8] P. LANCASTER AND M. TISMENETSKY, *The Theory of Matrices*, 2nd ed., Academic Press, Orlando, New York, 1985.
[9] L. D. LANDAU AND E. M. LIFSHITZ, *Course of Theoretical Physics,* Vol. 2: *The Classical Theory of Fields*, 4th ed., Pergamon Press, Oxford, New York, 1975.
[10] F. ROHRLICH, *Classical Charged Particles*, 2nd ed., Addison-Wesley, Reading, MA, 1990.
[11] K. SAKAMOTO, *Invariant manifolds in singular perturbation problems for ordinary differential equations*, Proc. Roy. Soc. Edinburgh Sect. A, 116 (1990), pp. 45–78.
[12] A. SOFFER AND M. WEINSTEIN, *Time dependent resonance theory*, Geom. Funct. Anal., 8 (1998), pp. 1086–1128.

[13] A. Soffer and M. Weinstein, *Resonances, radiation damping and instability in Hamiltonian nonlinear wave equations*, Invent. Math., 136 (1999), pp. 9–74.
[14] H. Spohn, *Runaway Charged Particles and Center Manifolds*, preprint, 1998.
[15] W. Thirring, *A Course in Mathematical Physics,* Vol. 2: *Classical Field Theory*, Springer-Verlag, New York, Vienna, 1978.
[16] A. D. Yaghjian, *Relativistic dynamics of a charged sphere*, Lecture Notes in Phys. 11, Springer-Verlag, Berlin, New York, 1992.

SIAM J. MATH. ANAL.
Vol. 32, No. 1, pp. 54–79

CLOSURE OF THE MANAKOV SYSTEM*

R. G. DOCKSEY† AND J. N. ELGIN†

Abstract. We discuss the closure of the product eigenstate for the Manakov system. Although the analysis is similar in principle to that for the Zakharov–Shabat system published elsewhere, two additional features arise which require careful attention.

First, in addition to the direct (or forward) scattering problem, an adjoint scattering problem is necessary. In the Zakharov–Shabat system, the adjoint problem is trivially related to the direct problem, leading to the formation of "squared" eigenstates rather than the product eigenstates which we derive.

Second, the system is not diagonal in the sense that both parts of the potential q_1 and q_2 contribute to each element S_{ij} of the scattering data, while—reciprocally—all (relevant) elements of the scattering data contribute in the reconstruction of both q_1 and q_2.

Key words. completeness, nonlinear Schrödinger equation, Manakov system

AMS subject classifications. 35Q55, 35P10, 58F19, 35Q60, 35Q51

PII. S0036141098343677

1. Introduction.

The vector nonlinear Schrödinger (NLS) equation is

$$i\mathbf{q_x} - \mathbf{q_{tt}} - \mathbf{2qq^\dagger q} = \mathbf{0}, \tag{1}$$

where a suffix denotes a partial derivative and † denotes Hermitian conjugation [1]. The most useful application of (1) is to studies on ultrashort pulse propagation down birefringent optical fibers [2], [3], [4], [5], [6]. In keeping with the notation now used in such studies, the roles of the independent variables x and t are such that *time* x is the distance propagated by the pulse down the fiber, while *spatial coordinate* t is a retarded time variable indicating position along the pulse. The Cauchy problem associated with (1) corresponds to specifying $\mathbf{q}(0,t)$, i.e., the profile of the pulse at input to the fiber. The vector nature of the dependant variable $\mathbf{q} = (q_1, q_2)^T$—where T denotes the transpose—is directly linked with the polarization state of the optical pulse.

Equation (1) is the lowest order nontrivial amplitude equation obtained from a multiple scales analysis of Maxwell's equations as appropriate to the fiber optic problem [2]. Both dispersion and nonlinearity are present, where the latter is the Kerr nonlinearity which corresponds to the intensity dependence of the refractive index of the host medium. The scalar form of (1) is obtained from the simple replacement of $\mathbf{q}$ with q and $\mathbf{q}^\dagger$ with q^*, where $*$ denotes complex conjugate; we will refer to this reduced form of (1) as the scalar problem.

Manakov [1] first demonstrated integrability of the vector NLS equation using the techniques of inverse scattering theory [7]. The Lax pair for the vector problem is a direct extension of that for the scalar problem first quoted by Zakharov and Shabat [8]. Kaup [9] investigated the closure of the squared Zakharov–Shabat eigenstates and motivated his work with the observation that it is precisely these squared eigenstates (and their closure) which are intrinsic to relating small changes in the potential q

*Received by the editors August 14, 1998; accepted for publication (in revised form) October 26, 1999; published electronically June 22, 2000.

http://www.siam.org/journals/sima/32-1/34367.html

†Department of Mathematics, Imperial College of Science, Technology and Medicine, London SW7 2BZ, UK (richard.docksey@barclayscapital.com, j.elgin@ic.ac.uk).

to small changes in the set of scattering data associated with the Zakharov–Shabat system, and vice versa. The corresponding statement for the Manakov system is made in (2) and (3) below.

In this article we extend Kaup's analysis to derive the closure for a set of product eigenstates associated with the Manakov system. In constructing this closure we need to take into account two features which do not arise in the scalar problem: First, an adjoint scattering problem is required in addition to the direct (or forward) scattering problem. With the Zakharov–Shabat system the adjoint problem is trivially related to the direct problem, leading to formation of squared eigenstates rather than the product states encountered here. Second, the system is not diagonal in the sense that both q_1 and q_2 contribute to each element S_{ij} of the scattering data, while—reciprocally—all (relevant) elements of the scattering data contribute in the reconstruction of both q_1 and q_2.

The spectral transform is a mapping from a potential $\mathbf{q}(x,t)$ into a set of scattering data $S_{ij}(\zeta,x)$ $(i,j=1,2,3)$, where ζ is an eigenparameter. The inverse map permits construction of the potential $\mathbf{q}$ from the set S_{ij}. Formally, we have (as will be proven in the later sections)

$$S_{ij}=\int_{-\infty}^{\infty}\phi^{(i)}\wedge\hat{\psi}^{(j)}\begin{pmatrix}\mathbf{q},\\ -\mathbf{q}^*\end{pmatrix}dt, \tag{2}$$

$$\begin{aligned}\begin{pmatrix}\mathbf{q}\\ \mathbf{q}^*\end{pmatrix}&=\frac{1}{\pi}\int_c\left(\frac{S_{21}}{S_{11}}\psi^{(2)}\vee\hat{\psi}^{(1)}+\frac{S_{31}}{S_{11}}\psi^{(3)}\vee\hat{\psi}^{(1)}\right)d\zeta\\ &\quad-\frac{1}{\pi}\int_{\bar{c}}\left(\frac{\Delta_{21}}{\Delta_{11}}\psi^{(1)}\vee\hat{\psi}^{(2)}+\frac{\Delta_{31}}{\Delta_{11}}\psi^{(1)}\vee\hat{\psi}^{(3)}\right)d\zeta.\end{aligned} \tag{3}$$

Here, $\phi^{(i)}\wedge\hat{\psi}^{(j)}$ and $\psi^{(i)}\vee\hat{\psi}^{(j)}$ are four component row and column vectors, respectively, whose components are made up of products between Jost function components for the forward and adjoint scattering problems (as discussed in section 3). It is the closure of those product states that is the main concern in this article.

The quantities Δ_{ij} are cofactors of S_{ij}, while the contours c and $\bar{c}$ are discussed later. For the moment, only one feature of (2) and (3) need be emphasized; namely, that the product states are intrinsic to the relationship between the potential $\mathbf{q}$ and the scattering data S_{ij}. A small change in $\mathbf{q}$ results in a change in S_{ij}, and vice versa.

Now suppose that the potential $\mathbf{q}(x,t)$ at time x changes to $\mathbf{q}(x+\delta x,t)$ at a later time $x+\delta x$ in accordance with (1). Then it follows from (2) that

$$\frac{d}{dx}S_{ij}=\int_{-\infty}^{\infty}\phi^{(i)}\wedge\hat{\psi}^{(j)}\begin{pmatrix}\mathbf{q}_x\\ -\mathbf{q}_x^*\end{pmatrix}dt. \tag{4}$$

For $(i,j)=(1,1)$, it can be shown independently that dS_{11}/dx is zero, so that $(\mathbf{q}_x,-\mathbf{q}_x^*)$ is orthogonal to the basis $\phi^{(i)}\wedge\hat{\psi}^{(j)}(\zeta;x,t)$. In other words, these product states are necessarily incomplete.

It is our intention here to find the closure of the product states of the Manakov system, and hence derive the completeness statement equation (78).

The article is arranged as follows: In section 2, we discuss the direct and adjoint scattering problems, and introduce the Jost function solutions of these equations, together with the scattering data. In the next section we introduce the product eigenstates—four component vectors which are bilinear products of an element from the Jost functions in each of the direct and adjoint scattering problems—and discuss

their closure. The statement of closure made in section 3 is proved in section 4 by redress to a set of Marchenko equations.

Some final comments are made in section 5; in particular, the completeness relation is used to express **q** in terms of integrals over the product states ((3) above), and we show how the family of vector NLS equations are generated using an integro-differential operator, introduced in section 3. The consequences of adding additional terms to (1) is then examined. Such terms may model the weak and strong birefringence properties of the fiber, the effects of higher order dispersion, loss in the fiber, and so on. Their effects on soliton propagation is examined: the vector soliton is a localized pulse which propagates without change of shape or polarization state. When any—or all—of these additional terms are included, the initial soliton state becomes modified in two ways: first, the soliton parameters and the polarization state of the pulse may change adiabatically with distance down the fiber. Second, the soliton may now shed radiation. Equation (3) is required to investigate the latter, which in turn follows from the closure of the product states. This application was the initial motivation behind this study.

2. The scattering problem. The scattering problem associated with (1) is [1], [10]

$$v_t + i\zeta E v = Q v, \tag{5}$$

where $v = [v_1, v_2, v_3]^T$ is a spinor eigenstate, (complex) ζ is an eigenparameter, the suffix t denotes the derivative $\partial/\partial t$, and the matrices E and Q are defined by

$$E = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix}, \qquad Q = \begin{pmatrix} 0 & q_1 & q_2 \\ -q_1^* & 0 & 0 \\ -q_2^* & 0 & 0 \end{pmatrix}. \tag{6}$$

In this article, (5) is considered on the infinite interval $-\infty < t < \infty$, while the potentials q_1 and $q_2 \in L_1$, i.e., they satisfy

$$\begin{aligned} &\int_{-\infty}^{\infty} |q_1(x,t)|dt < \infty, \\ &\int_{-\infty}^{\infty} |q_2(x,t)|dt < \infty. \end{aligned} \tag{7}$$

We define the fundamental (or Jost) solutions $\phi^{(i)}$ and $\psi^{(i)}$, and $i = 1, 2, 3$, for real $\zeta = \xi$ by the requirements that

$$\phi^{(1)} \sim \begin{pmatrix} e^{-i\xi t} \\ 0 \\ 0 \end{pmatrix}, \quad \phi^{(2)} \sim \begin{pmatrix} 0 \\ e^{i\xi t} \\ 0 \end{pmatrix}, \quad \phi^{(3)} \sim \begin{pmatrix} 0 \\ 0 \\ e^{i\xi t} \end{pmatrix}, \quad t \to -\infty,$$

$$\psi^{(1)} \sim \begin{pmatrix} e^{-i\xi t} \\ 0 \\ 0 \end{pmatrix}, \quad \psi^{(2)} \sim \begin{pmatrix} 0 \\ e^{i\xi t} \\ 0 \end{pmatrix}, \quad \psi^{(3)} \sim \begin{pmatrix} 0 \\ 0 \\ e^{i\xi t} \end{pmatrix}, \quad t \to \infty.$$

Throughout this article, a superscript will be used to denote one of the Jost function solutions, and a subscript will denote a particular component of that solution, so that $\phi_1^{(2)}$ is the first component of the Jost function, $\phi^{(2)}$.

Since $\phi^{(i)}$ and $\psi^{(j)}$ are independent sets of solutions, we can write

$$\phi^{(i)} = \sum_{j=1}^{3} S_{ji}(\zeta)\psi^{(j)}, \tag{8}$$

which defines the scattering data $S_{ij}(\zeta)$. For $\zeta = \xi$ real, S is a 3×3 unitary unimodular matrix [10].

We also require an adjoint scattering problem which is taken to be

$$\hat{v}_t - i\zeta E\hat{v} = Q^*\hat{v}, \tag{9}$$

where the $\hat{}$ symbol will be used to denote solutions of the adjoint problem. As with the direct problem, we define the fundamental solutions $\hat{\phi}^{(i)}$ and $\hat{\psi}^{(i)}$ of the adjoint problem by the requirement that

$$\hat{\phi}^{(1)} \sim \begin{pmatrix} e^{i\xi t} \\ 0 \\ 0 \end{pmatrix}, \quad \hat{\phi}^{(2)} \sim \begin{pmatrix} 0 \\ e^{-i\xi t} \\ 0 \end{pmatrix}, \quad \hat{\phi}^{(3)} \sim \begin{pmatrix} 0 \\ 0 \\ e^{-i\xi t} \end{pmatrix}, \quad t \to -\infty,$$

$$\hat{\psi}^{(1)} \sim \begin{pmatrix} e^{i\xi t} \\ 0 \\ 0 \end{pmatrix}, \quad \hat{\psi}^{(2)} \sim \begin{pmatrix} 0 \\ e^{-i\xi t} \\ 0 \end{pmatrix}, \quad \hat{\psi}^{(3)} \sim \begin{pmatrix} 0 \\ 0 \\ e^{-i\xi t} \end{pmatrix}, \quad t \to \infty.$$

Note that for the scattering problem associated with the scalar NLS equation, the two component adjoint Jost functions are related in a simple way to the direct functions by $\hat{\phi}^{(1)} = \sigma_1\phi^{(2)}$, $\hat{\phi}^{(2)} = \sigma_1\phi^{(1)}$, where σ_1 is the Pauli matrix. In view of this, no explicit statement of an adjoint problem is required.

Since by construction $(\hat{\psi}^{(i)})^T\psi^{(j)} = \delta_{ij}$, it follows that $S_{ij} = (\hat{\psi}^{(i)})^T\phi^{(j)}$. The scattering data Δ_{ij} for the adjoint scattering problem are introduced in a manner analogous to (8) by

$$\hat{\phi}^{(i)} = \sum_{j=1}^{3} \Delta_{ji}(\zeta)\hat{\psi}^{(j)}. \tag{10}$$

By virtue of the unitary nature of S, it is easily demonstrated that $\Delta_{ji}(\zeta)$ is the cofactor of the element $S_{ij}(\zeta)$ and that

$$\Delta_{ij}(\zeta) = S_{ij}^*(\zeta), \tag{11}$$

where $*$ denotes complex conjugate.

A simple consideration of the analytic properties of $\phi^{(1)}(\zeta)$, etc., reveals the following [10]:

- $\phi^{(1)}, \psi^{(2)}, \psi^{(3)}, \hat{\phi}^{(2)}, \hat{\phi}^{(3)}, \hat{\psi}^{(1)}$ are analytic in the upper half of the complex ζ-plane, as are $S_{11}, \Delta_{22}, \Delta_{33}, \Delta_{23}$, and Δ_{32};
- $\phi^{(2)}, \phi^{(3)}, \psi^{(1)}, \hat{\phi}^{(1)}, \hat{\psi}^{(2)}, \hat{\psi}^{(3)}$ are analytic in the lower half of the complex ζ-plane, as are $S_{22}, S_{33}, S_{23}, S_{32}$, and Δ_{11};
- $S_{12}, S_{13}, S_{21}, S_{31}, \Delta_{12}, \Delta_{13}, \Delta_{21}, \Delta_{31}$ are defined only for $\zeta = \xi$ real, with possible continuation into the ζ-plane dependant on the manner in which $|q_1|$ and $|q_2|$ decrease as $|t| \to \infty$. For example, if $q_1,\ q_2 < Ce^{-2K|t|}$ as $t \to \infty$, these quantities are analytic in the strip $K > \Im\zeta > -K$.

The zeros of $S_{11}(\zeta)$ in the upper half plane, $S_{11}(\zeta_k) = 0$, $k = 1, \ldots, N$, give the bound states of (5). We will assume throughout that such zeros are simple, though our results will continue to hold for cases where this is not so. With $S_{11}(\zeta_k) = 0$, (8) gives

$$\phi^{(1)}(\zeta_k, t) = c_k \psi^{(2)}(\zeta_k, t) + d_k \psi^{(3)}(\zeta_k, t), \tag{12}$$

with suitably chosen weights c_k and d_k. To construct similar bound state eigenfunctions $\phi^{(2)}$ and $\phi^{(3)}$, it would appear that we require the (overrestrictive) conditions that S_{22}, S_{23}, S_{32}, and S_{33} all vanish at some $\zeta = \bar{\zeta}_k$, $\Im\bar{\zeta}_k < 0$ in order to eliminate the exponentially growing terms at $t \to \infty$. These requirements can be relaxed by introducing a new pair of eigenfunctions $\phi^{(\pm)}$, analytic in the lower half plane, to replace $\phi^{(2)}$ and $\phi^{(3)}$. Define

$$\phi^{(+)}(\zeta, t) = S_{33}(\zeta)\phi^{(2)}(\zeta, t) - S_{32}(\zeta)\phi^{(3)}(\zeta, t) \tag{13}$$

$$= \Delta_{11}(\zeta)\psi^{(2)}(\zeta, t) - \Delta_{21}(\zeta)\psi^{(1)}(\zeta, t), \tag{14}$$

$$\phi^{(-)}(\zeta, t) = S_{22}(\zeta)\phi^{(3)}(\zeta, t) - S_{23}(\zeta)\phi^{(2)}(\zeta, t) \tag{15}$$

$$= \Delta_{11}(\zeta)\psi^{(3)}(\zeta, t) - \Delta_{31}(\zeta)\psi^{(1)}(\zeta, t), \tag{16}$$

where the second forms of each set follow from using (8).

At $\bar{\zeta}_k = \zeta_k^*$, $\Delta_{11}(\bar{\zeta}_k) = 0$, and hence we have the bound states

$$\phi^{(+)}(\bar{\zeta}_k, t) = \bar{c}_k \psi^{(1)}(\bar{\zeta}_k, t), \tag{17}$$

$$\phi^{(-)}(\bar{\zeta}_k, t) = \bar{d}_k \psi^{(1)}(\bar{\zeta}_k, t), \tag{18}$$

with suitably chosen weights $\bar{c}_k$ and $\bar{d}_k$. In Appendix A, similar definitions are given for the functions $\psi^{(\pm)}$, $\hat{\phi}^{(\pm)}$, and $\hat{\psi}^{(\pm)}$.

3. The product Manakov eigenstates and their closure. In this section we will discuss the product eigenstates of the Manakov system. These are four component vectors whose components are bilinear products of an element from each of the direct and adjoint scattering problems, such as $\psi_k^{(i)}\hat{\psi}_l^{(j)}$ for some i, j, k, l. From the scattering problem and its adjoint, we find that such product states satisfy a ninth order system of equations, which can be reduced to a smaller set of four equations by formal integration of five of the set. This procedure results in the introduction of a 4×4 integro-differential matrix operator $\mathcal{L}$,

$$\mathcal{L} = \frac{1}{2i} \begin{pmatrix} -\partial_t + J^{(1)}_{\mathbf{q}\mathbf{q}^*} & -J^{(2)}_{\mathbf{q}\mathbf{q}} \\ J^{(2)}_{\mathbf{q}^*\mathbf{q}^*} & \partial_t - J^{(1)}_{\mathbf{q}^*\mathbf{q}} \end{pmatrix}. \tag{19}$$

The integral operators $J^{(1)}$ and $J^{(2)}$ are defined by their action on an arbitrary vector function $\mathbf{f}$,

$$J^{(1)}_{\mathbf{u}\mathbf{v}}[\mathbf{f}] = \int_t^\infty \left(\mathbf{u}(t)\mathbf{v}^T(t') + \mathbf{v}^T(t')\mathbf{u}(t)\right)\mathbf{f}(t')dt', \tag{20}$$

$$J^{(2)}_{\mathbf{u}\mathbf{v}}[\mathbf{f}] = \int_t^\infty \left(\mathbf{u}(t)\mathbf{v}^T(t') + \mathbf{v}(t')\mathbf{u}^T(t)\right)\mathbf{f}(t')dt', \tag{21}$$

where $\mathbf{u}$ and $\mathbf{v}$ are (arbitrary) vector suffixes. The suffixes $\mathbf{q}$ and $\mathbf{q}^*$ which appear in (19) correspond to the potential $\mathbf{q}$ and its conjugate $\mathbf{q}^*$.

On reducing the ninth order system to a set of four coupled equations, the remaining set of four product variables can be written as the components of a four vector, which we now introduce with the following notation. For any u and $\hat{v}$, three component vector solutions of the scattering problem and its adjoint, respectively, define

$$u \vee \hat{v} = (u_1\hat{v}_2, u_1\hat{v}_3, -u_2\hat{v}_1, -u_3\hat{v}_1)^T. \tag{22}$$

Hence, for example,

$$\psi^{(i)} \vee \hat{\psi}^{(j)} = \left(\psi_1^{(i)}\hat{\psi}_2^{(j)}, \psi_1^{(i)}\hat{\psi}_3^{(j)}, -\psi_2^{(i)}\hat{\psi}_1^{(j)}, -\psi_3^{(i)}\hat{\psi}_1^{(j)}\right)^T. \tag{23}$$

One can easily show that $\psi^{(1)} \vee \hat{\psi}^{(2)}$, $\psi^{(1)} \vee \hat{\psi}^{(3)}$, $\psi^{(2)} \vee \hat{\psi}^{(1)}$, and $\psi^{(3)} \vee \hat{\psi}^{(1)}$ are all eigenfunctions of the operator $\mathcal{L}$, with eigenvalue ζ, i.e.,

$$\left.\begin{array}{l} \mathcal{L}\psi^{(1)} \vee \hat{\psi}^{(2)} = \zeta\psi^{(1)} \vee \hat{\psi}^{(2)} \\ \mathcal{L}\psi^{(1)} \vee \hat{\psi}^{(3)} = \zeta\psi^{(1)} \vee \hat{\psi}^{(3)} \end{array}\right\} \Im\zeta \le 0, \tag{24}$$

and

$$\left.\begin{array}{l} \mathcal{L}\psi^{(2)} \vee \hat{\psi}^{(1)} = \zeta\psi^{(2)} \vee \hat{\psi}^{(1)} \\ \mathcal{L}\psi^{(3)} \vee \hat{\psi}^{(1)} = \zeta\psi^{(3)} \vee \hat{\psi}^{(1)} \end{array}\right\} \Im\zeta \ge 0. \tag{25}$$

Other choices of i, j produce product states which are not eigenstates of $\mathcal{L}$; we will refer to these states as "cross states." These satisfy the relationship

$$\mathcal{L}\psi^{(i)} \vee \hat{\psi}^{(j)} = \zeta\psi^{(i)} \vee \hat{\psi}^{(j)} - \frac{1}{2i}\gamma_{ij}\begin{pmatrix} \mathbf{q} \\ \mathbf{q}^* \end{pmatrix}, \tag{26}$$

where the matrices γ_{ij} are given by

$$\begin{aligned} \gamma_{11} &= -\mathbf{I}_4, \\ \gamma_{22} &= \tfrac{1}{2}\begin{pmatrix} \mathbf{I}_2 + \sigma_3 & 0 \\ 0 & \mathbf{I}_2 + \sigma_3 \end{pmatrix}, \\ \gamma_{23} &= \begin{pmatrix} \sigma_- & 0 \\ 0 & \sigma_+ \end{pmatrix}, \\ \gamma_{32} &= \begin{pmatrix} \sigma_+ & 0 \\ 0 & \sigma_- \end{pmatrix}, \\ \gamma_{33} &= \tfrac{1}{2}\begin{pmatrix} \mathbf{I}_2 - \sigma_3 & 0 \\ 0 & \mathbf{I}_2 - \sigma_3 \end{pmatrix}, \end{aligned} \tag{27}$$

with

$$\sigma_+ = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \qquad \sigma_- = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \qquad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{28}$$

The remaining $\gamma_{12}, \gamma_{13}, \gamma_{21}$, and γ_{31} matrices are zero, leading to (24) and (25).

Vectors such as $\psi^{(1)} \vee \hat{\psi}^{(2)}$ will form the basis states in our vector space. To define and evaluate inner products between such vectors, we will require an adjoint problem associated with (26) above. The adjoint operator $\mathcal{L}_A$ is now defined to be

$$\mathcal{L}_A = \frac{1}{2i} \begin{pmatrix} \partial_t + I^{(1)}_{\mathbf{q}^*\mathbf{q}} & I^{(2)}_{\mathbf{q}^*\mathbf{q}^*} \\ -I^{(2)}_{\mathbf{qq}} & -\partial_t - I^{(1)}_{\mathbf{qq}^*} \end{pmatrix}, \tag{29}$$

where the integral operators $I^{(1)}_{\mathbf{uv}}$ and $I^{(2)}_{\mathbf{uv}}$ are defined by the actions

$$I^{(1)}_{\mathbf{uv}}[\mathbf{f}] = \int_{-\infty}^{t} \left(\mathbf{u}(t)\mathbf{v}^T(t') + \mathbf{v}^T(t')\mathbf{u}(t)\right) \mathbf{f}(t')dt', \tag{30}$$

$$I^{(2)}_{\mathbf{uv}}[\mathbf{f}] = \int_{-\infty}^{t} \left(\mathbf{u}(t)\mathbf{v}^T(t') + \mathbf{v}(t')\mathbf{u}^T(t)\right) \mathbf{f}(t')dt'. \tag{31}$$

Here $\mathbf{u}$ and $\mathbf{v}$ are (arbitrary) vector suffixes, and $\mathbf{f}$ is an arbitrary vector function as before.

By analogy with (22), the adjoint product states are introduced as

$$u \wedge \hat{v} = (u_2\hat{v}_1, u_3\hat{v}_1, u_1\hat{v}_2, u_1\hat{v}_3) \tag{32}$$

so that, for example,

$$\phi^{(i)} \wedge \hat{\phi}^{(j)} = \left(\phi^{(i)}_2\hat{\phi}^{(j)}_1, \phi^{(i)}_3\hat{\phi}^{(j)}_1, \phi^{(i)}_1\hat{\phi}^{(j)}_2, \phi^{(i)}_1\hat{\phi}^{(j)}_3\right). \tag{33}$$

Note that $u \wedge \hat{v}$ is to be interpreted as a row vector, whereas $u \vee \hat{v}$ is a column vector.

It can be shown that $\phi^{(1)} \wedge \hat{\phi}^{(2)}$, $\phi^{(1)} \wedge \hat{\phi}^{(3)}$, $\phi^{(2)} \wedge \hat{\phi}^{(1)}$, and $\phi^{(3)} \wedge \hat{\phi}^{(1)}$ are eigenstates of $\mathcal{L}_A$, while the cross states satisfy

$$\mathcal{L}_A \left(\phi^{(i)} \wedge \hat{\phi}^{(j)}\right)^T = \zeta \left(\phi^{(i)} \wedge \hat{\phi}^{(j)}\right)^T + \frac{1}{2i}\gamma_{ji} \begin{pmatrix} \mathbf{q}^* \\ -\mathbf{q} \end{pmatrix}, \tag{34}$$

where the matrices γ_{ji} are as defined previously. Note the transposition of the suffixes in γ_{ij} in (34).

We will now proceed to define an inner product between the product states and the adjoint product states. It will be shown that the inner product between any bound product eigenstate (such as $\psi^{(1)} \vee \hat{\psi}^{(2)}$ evaluated at $\zeta = \bar{\zeta}_k = \zeta_k^*$) with any other eigenstate vanishes. Consequently, when the scattering problem supports bound eigenstates, the product eigenstates cannot be complete. As with the Zakharov–Shabat scattering problem, the remedy to finding closure is with the cross terms. We show that the inner product between any bound eigenstate with a cross term is nonzero. This leads us to construct a basis consisting of the product eigenstates together with the cross terms. Investigation of the linear independence of this basis reveals that the cross terms are linearly independent only when evaluated at the bound state eigenvalues, and that the linearly independent component of the cross states is the derivative of the product eigenstates with respect to the eigenvalue parameter, evaluated at the bound state eigenvalue. Following Kaup, we will refer to these derivative states as "P-states." The new basis is then taken to be the product eigenstates, together with the P-states. Finally, we construct a resolution of the identity operator.

A formal proof of completeness is deferred until the next section.

To simplify notation, we will use the Dirac bra- and ket- notation. Let $|t\rangle$ and $\langle t|$ be complete sets of eigenstates of the position vector, and define

$$\langle t|ij,\zeta\rangle = \psi^{(i)} \vee \hat{\psi}^{(j)}(\zeta,t), \tag{35}$$

$$\langle ij,\zeta|t\rangle = \phi^{(i)} \wedge \hat{\phi}^{(j)}(\zeta,t) \tag{36}$$

for $\zeta = \xi$ real and for continuation into the appropriate half space. Note that, consistent with our previous definitions, we interpret $\psi^{(i)} \vee \hat{\psi}^{(j)}(\zeta,t)$ as a column vector and $\phi^{(i)} \wedge \hat{\phi}^{(j)}(\zeta,t)$ as a row vector.

Suppose now that U^T is an arbitrary row vector and V is an arbitrary column vector. Then, the formal adjoint operator $\mathcal{L}_A$ was constructed such that the following relationship is true:

$$U^T\mathcal{L}V = V^T\mathcal{L}_A U - \frac{1}{2i}\partial_t f \tag{37}$$

with $f = f_1 - f_2$, where

$$f_1 = V^T\Sigma_3 U, \tag{38}$$

$$f_2 = \left\{ I_+ U^T \begin{pmatrix} \mathbf{q}^* \\ -\mathbf{q} \end{pmatrix} \right\} \cdot \left\{ I_- V^T \begin{pmatrix} \mathbf{q} \\ \mathbf{q}^* \end{pmatrix} \right\} + I_+ \left(U^T I_-(MV) \right). \tag{39}$$

Here,

$$\Sigma_3 = \begin{pmatrix} \mathbf{I}_2 & 0 \\ 0 & -\mathbf{I}_2 \end{pmatrix} \tag{40}$$

and M is the 4×4 matrix

$$M = \begin{pmatrix} \mathbf{q}^\dagger\mathbf{q}\mathbf{I}_2 & \mathbf{q}^*\mathbf{q}^\dagger \\ -\mathbf{q}\mathbf{q}^T & -\mathbf{q}^\dagger\mathbf{q}\mathbf{I}_2 \end{pmatrix}. \tag{41}$$

The integral operator $I_+(I_-)$ denotes integration from t to $+\infty$ ($-\infty$ to t), and it is understood that in matrix M, the first $\mathbf{q}$ variable is to appear in the integral I_+ and the second in I_- (so that, if t' and t'' are the integration variables in I_+ and I_-, respectively, $\mathbf{q}^\dagger\mathbf{q}\mathbf{I}_2 = \mathbf{q}^\dagger(t')\mathbf{q}(t'')\mathbf{I}_2$). For the simpler Zakharov–Shabat system, $\mathbf{q}$ becomes scalar q and the matrix M simplifies to

$$M = \begin{pmatrix} q^*(t') \\ -q(t') \end{pmatrix} \cdot \left(q(t''), q^*(t'') \right) \tag{42}$$

so that the second contribution to f_2 is then the same as the first, leading to results published by Kaup [9].

Note that, by virtue of the imposed conditions, (7), f_2 will vanish as $t \to \pm\infty$, while f_1 may remain finite, depending on choices for the vectors U and V.

In terms of Dirac notation, (26) and (34) read

$$\begin{aligned} \mathcal{L}|ij,\zeta\rangle &= \zeta|ij,\zeta\rangle - \frac{1}{2i}\gamma_{ij}|p\rangle, \\ \langle ij,\zeta|\mathcal{L}_A &= \zeta\langle ij,\zeta| + \frac{1}{2i}\gamma_{ji}\langle p|, \end{aligned} \tag{43}$$

where the vectors $|p\rangle$ and $\langle p|$ are defined by

$$\langle t|p\rangle = \begin{pmatrix} \mathbf{q}(t,x) \\ \mathbf{q}^*(t,x) \end{pmatrix}, \tag{44}$$

$$\langle p|t\rangle = \begin{pmatrix} \mathbf{q}^*(t,x) \\ -\mathbf{q}(t,x) \end{pmatrix}, \tag{45}$$

and it is understood that $\mathcal{L}_A$ always operates backwards on the preceeding bra. We are now in a position to define and evaluate inner products.

We define the inner product between product states by

$$\langle ij,\zeta'|kl,\zeta\rangle = \lim_{T\to\infty}\int_{-T}^{T}\left(\phi^{(i)}\wedge\hat{\phi}^{(j)}(\zeta',t)\right)\psi^{(k)}\vee\hat{\psi}^{(l)}(\zeta,t)dt. \tag{46}$$

To evaluate the integrals, and hence find the inner products, we use (43) and (37) together with the definitions of the Jost functions, and the scattering data equations (8) and (10) inserted as appropriate into the expression for f_1, (38), taken in the limit $t\to\pm\infty$. When $\zeta=\xi$ real, it is then straightforward to derive the inner products for the continuum states; these are quoted in Table 1 (we do not show results for all inner products since not all are necessary for our proof of completeness). For the bound product states, the inner products between eigenstates of the operators $\mathcal{L}$ and $\mathcal{L}_A$ are shown in Table 2. Contrary to corresponding results quoted for the Zakharov–Shabat system, these are not all equal to zero. However, the results quoted in Table 2 make use of eigenstates like $\phi^{(2)}$ and $\phi^{(3)}$ in the construction of product states and, as discussed previously, these are not suitable candidates for the bound state solutions of (5); rather, one should use $\phi^{(+)}$ and $\phi^{(-)}$. Hence, we now introduce a new set of product states $\psi^{(1)}\vee\hat{\psi}^{(+)}$, etc., defined as before in (23), but now the superfix $\pm$ replaces $j=2,3$ as appropriate. Matrix elements between members of this new set of product eigenstates can then be evaluated readily from results quoted in Table 2. All such inner products are found to be identically zero, i.e., $\langle 1+,\zeta_k|1+,\bar{\zeta}_l\rangle = \langle 1+,\zeta_k|+1,\zeta_l\rangle = \langle +1,\bar{\zeta}_k|1-,\bar{\zeta}_l\rangle = \cdots = 0$. The reason for this vanishing of the bound state inner products is made apparent by returning to the results for the continuum states in Table 1; in the new basis, the inner products are as listed in Table 3. All such products between eigenstates contain either S_{11}^2 or Δ_{11}^2, corresponding to double zeros (since S_{11} and Δ_{11} have simple zeros) on continuation away from the real ζ axis to a zero of S_{11} (or Δ_{11}).

The inner product between any bound state and any continuum state is identically zero.

Clearly, the new set of product eigenstates is incomplete. To see how to complete them we return again to the observation that all matrix elements between continuum product eigenstates are proportional either to $S_{11}^2(\xi)$ or $\Delta_{11}^2(\xi)$ $(=S_{11}^*(\xi)^2)$. As we will show, this requires that the continuum contribution to the completeness statement is an integral where coefficients appear which are proportional to $S_{11}^{-2}(\xi)$ (or $\Delta_{11}^{-2}(\xi)$). If, for temporary convenience, an assumption is made that the potentials q_1 and q_2 have compact support, then extending this integral contribution into the complex plane would result in a contribution from the double poles of $S_{11}^{-2}(\zeta)$ at $\zeta=\zeta_k$ or from $\Delta_{11}^{-2}(\zeta)$ at $\zeta=\bar{\zeta}_k$. Consequently, the residual contributions from these poles will result not only in derivatives (with respect to ζ) of other coefficients, but also derivatives of the bra- and ket-vectors which appear in the completeness statement. Such vectors are the product eigenstates $|1+,\zeta\rangle$, etc., of $\mathcal{L}$ (similarly, $\mathcal{L}_A$) discussed above.

TABLE 1

Inner products between continuum product eigenstates. Here $\alpha = \pi\delta(\xi - \xi')$. *Also shown are matrix elements between eigenstates and the product cross states* $|11,\xi\rangle$ *and* $\langle 11,\xi'|$. *Here,* $\beta = i/(2(\xi' - \xi + i\epsilon))$ *taken in the limit* $\epsilon \to 0$. *Note that all the matrix elements are functions of* ξ *except in the last column where they are functions of* ξ'.

	$\|12,\xi\rangle$	$\|13,\xi\rangle$	$\|21,\xi\rangle$	$\|31,\xi\rangle$	$\|11,\xi\rangle$
$\langle 12,\xi'\|$	0	0	$-\alpha S_{11}\Delta_{22}$	$-\alpha S_{11}\Delta_{32}$	$-\beta S_{11}\Delta_{12}$
$\langle 13,\xi'\|$	0	0	$-\alpha S_{11}\Delta_{23}$	$-\alpha S_{11}\Delta_{33}$	$-\beta S_{11}\Delta_{13}$
$\langle 21,\xi'\|$	$\alpha\Delta_{11}S_{22}$	$\alpha\Delta_{11}S_{32}$	0	0	$\beta^*\Delta_{11}S_{12}$
$\langle 31,\xi'\|$	$\alpha\Delta_{11}S_{23}$	$\alpha\Delta_{11}S_{33}$	0	0	$\beta^*\Delta_{11}S_{13}$
$\langle 11,\xi'\|$	$\beta\Delta_{11}S_{21}$	$\beta\Delta_{11}S_{31}$	$-\beta^*S_{11}\Delta_{21}$	$-\beta^*S_{11}\Delta_{31}$	0

TABLE 2

Inner products between bound product eigenstates. The elements in the first two columns are functions of $\bar{\zeta}_k$ *and those in the last two columns are functions of* ζ. *Here,* S'_{11} *means* $dS_{11}(\zeta)/d\zeta$ *evaluated at* $\zeta = \zeta_k$, *and similarly for* $\Delta'_{11}(\bar{\zeta}_k)$

	$\|12,\bar{\zeta}_l\rangle$	$\|13,\bar{\zeta}_l\rangle$	$\|21,\bar{\zeta}_l\rangle$	$\|31,\bar{\zeta}_l\rangle$
$\langle 12,\zeta_k\|$	0	0	$\frac{1}{2i}S'_{11}\Delta_{22}\delta_{kl}$	$\frac{1}{2i}S'_{11}\Delta_{32}\delta_{kl}$
$\langle 13,\zeta_k\|$	0	0	$\frac{1}{2i}S'_{11}\Delta_{23}\delta_{kl}$	$\frac{1}{2i}S'_{11}\Delta_{33}\delta_{kl}$
$\langle 21,\bar{\zeta}_k\|$	$\frac{1}{2i}\Delta'_{11}S_{22}\delta_{kl}$	$\frac{1}{2i}\Delta'_{11}S_{32}\delta_{kl}$	0	0
$\langle 31,\bar{\zeta}_k\|$	$\frac{1}{2i}\Delta'_{11}S_{23}\delta_{kl}$	$\frac{1}{2i}\Delta'_{11}S_{33}\delta_{kl}$	0	0

This leads us to propose that the following set of states will complete our space:

$$|(+1)P,\zeta_k\rangle = \frac{\partial}{\partial\zeta}|+1,\zeta\rangle_{\zeta=\zeta_k}, \tag{47}$$

$$|(-1)P,\zeta_k\rangle = \frac{\partial}{\partial\zeta}|-1,\zeta\rangle_{\zeta=\zeta_k}, \tag{48}$$

$$|(1+)P,\bar{\zeta}_k\rangle = \frac{\partial}{\partial\zeta}|1+,\zeta\rangle_{\zeta=\bar{\zeta}_k}, \tag{49}$$

$$|(1-)P,\bar{\zeta}_k\rangle = \frac{\partial}{\partial\zeta}|1-,\zeta\rangle_{\zeta=\bar{\zeta}_k}. \tag{50}$$

A similar definition applies to the bra-states. We follow Kaup's terminology and refer to these as "P-states."

The P-states satisfy the equations

$$\mathcal{L}|(+1)P,\zeta_k\rangle = \zeta_k|(+1)P,\zeta_k\rangle + |+1,\zeta_k\rangle, \tag{51}$$

$$\langle(1+)P,\zeta_k|\mathcal{L}_A = \zeta_k\langle(1+)P,\zeta_k| + \langle 1+,\zeta_k|, \tag{52}$$

and similarly for the other states. Note that the P-states are not eigenstates of $\mathcal{L}$ (nor of $\mathcal{L}_A$). Equations (51) and (52) follow from straightforward differentiation of (24), (25), and (34) with respect to ζ.

To use the P-states, we need first to evaluate the inner products between these and the product eigenstates $|1+,\zeta\rangle$, etc., as well as between the P-states themselves. For inner products between bound states, the only nonzero inner products between a

TABLE 3

Inner products between dressed continuum eigenstates. Also shown are matrix elements between dressed eigenstates and the product cross states $|11,\xi\rangle$ and $\langle 11,\xi'|$. α and β are defined in the caption to Table 1. Note that all matrix elements in the last row are functions of ξ while all elements in the last column are functions of ξ'.

	$\|1+,\xi\rangle$	$\|1-,\xi\rangle$	$\|+1,\xi\rangle$	$\|-1,\xi\rangle$	$\|11,\xi\rangle$
$\langle 1+,\xi'\|$	0	0	$-\alpha S_{11}^2\Delta_{33}$	$\alpha S_{11}^2\Delta_{32}$	$\beta S_{11}S_{21}$
$\langle 1-,\xi'\|$	0	0	$\alpha S_{11}^2\Delta_{23}$	$-\alpha S_{11}^2\Delta_{22}$	$\beta S_{11}S_{31}$
$\langle +1,\xi'\|$	$\alpha\Delta_{11}^2 S_{33}$	$-\alpha\Delta_{11}^2 S_{32}$	0	0	$-\beta^*\Delta_{11}\Delta_{21}$
$\langle -1,\xi'\|$	$-\alpha\Delta_{11}^2 S_{23}$	$\alpha\Delta_{11}^2 S_{22}$	0	0	$-\beta^*\Delta_{11}\Delta_{31}$
$\langle 11,\xi'\|$	$-\beta\Delta_{11}\Delta_{12}$	$-\beta\Delta_{11}\Delta_{13}$	$\beta^* S_{11}S_{12}$	$\beta^* S_{11}S_{13}$	0

P-state and an eigenstate are

$$\langle (1+)P,\zeta_k|+1,\zeta_l\rangle = \langle 1+,\zeta_k|(+1)P,\zeta_l\rangle = -i\Delta_{33}(\zeta_k)\left(S'_{11}(\zeta_l)\right)^2\delta_{kl}, \tag{53}$$

$$\langle (1-)P,\zeta_k|-1,\zeta_l\rangle = \langle 1-,\zeta_k|(-1)P,\zeta_l\rangle = -i\Delta_{22}(\zeta_k)\left(S'_{11}(\zeta_l)\right)^2\delta_{kl}, \tag{54}$$

$$\langle (1+)P,\zeta_k|-1,\zeta_l\rangle = \langle 1+,\zeta_k|(-1)P,\zeta_l\rangle = i\Delta_{32}(\zeta_k)\left(S'_{11}(\zeta_l)\right)^2\delta_{kl}, \tag{55}$$

$$\langle (1-)P,\zeta_k|+1,\zeta_l\rangle = \langle 1-,\zeta_k|(+1)P,\zeta_l\rangle = i\Delta_{23}(\zeta_k)\left(S'_{11}(\zeta_l)\right)^2\delta_{kl}, \tag{56}$$

$$\langle (+1)P,\bar\zeta_k|1+,\bar\zeta_l\rangle = \langle +1,\bar\zeta_k|(1+)P,\bar\zeta_l\rangle = -iS_{33}(\bar\zeta_k)\left(\Delta'_{11}(\bar\zeta_l)\right)^2\delta_{kl}, \tag{57}$$

$$\langle (-1)P,\bar\zeta_k|1-,\bar\zeta_l\rangle = \langle -1,\bar\zeta_k|(1-)P,\bar\zeta_l\rangle = -iS_{22}(\bar\zeta_k)\left(\Delta'_{11}(\bar\zeta_l)\right)^2\delta_{kl}, \tag{58}$$

$$\langle (+1)P,\bar\zeta_k|1-,\bar\zeta_l\rangle = \langle +1,\bar\zeta_k|(1-)P,\bar\zeta_l\rangle = iS_{32}(\bar\zeta_k)\left(\Delta'_{11}(\bar\zeta_l)\right)^2\delta_{kl}, \tag{59}$$

$$\langle (-1)P,\bar\zeta_k|1+,\bar\zeta_l\rangle = \langle -1,\bar\zeta_k|(1+)P,\bar\zeta_l\rangle = iS_{23}(\bar\zeta_k)\left(\Delta'_{11}(\bar\zeta_l)\right)^2\delta_{kl}, \tag{60}$$

where $S'_{11}(\zeta_l)$ means $dS_{11}/d\zeta|_{\zeta=\zeta_l}$, and similarly for $\Delta'_{11}(\bar\zeta_l)$.

To complete the set, we require inner products between pairs of P-states. It is found that $\langle (1\mu)P,\zeta_k|(\nu 1)P,\zeta_l\rangle$, with μ and ν equal to $\pm$ in turn, is zero unless $k=l$. With $k=l$ it is found that

$$\langle (1\mu)P,\zeta_k|(\nu 1)P,\zeta_k\rangle = \frac{\partial}{\partial\zeta}\langle(\mu 1)P,\zeta|\nu 1,\zeta\rangle_{\zeta=\zeta_k} \tag{61}$$

and, similarly,

$$\langle (\mu 1)P,\bar\zeta_k|(1\nu)P,\bar\zeta_k\rangle = \frac{\partial}{\partial\zeta}\langle(1\mu)P,\zeta|1\nu,\zeta\rangle_{\zeta=\bar\zeta_k}. \tag{62}$$

So, for example,

$$\langle (1+)P,\zeta_k|(+1)P,\zeta_l\rangle = -i\left(\Delta_{33}(\zeta)S'_{11}(\zeta)^2\right)'_{\zeta=\zeta_k}\delta_{kl}. \tag{63}$$

We will demonstrate below that the full basis constructed from the continuum eigenstates, the bound eigenstates, and these P-states is complete with respect to the $L_2(-\infty,\infty)$ norm. In consequence, any ket $|u\rangle$, say, can be expressed in terms of this set of basis functions as

$$
\begin{aligned}
|u\rangle = \frac{1}{\pi}\int_{-\infty}^{\infty} d\xi \left(f(\xi)|+1,\xi\rangle + g(\xi)|-1,\xi\rangle + \bar{f}(\xi)|1+,\xi\rangle + \bar{g}(\xi)|1-,\xi\rangle \right) \\
+ \sum_{k=1}^{N} \left(f_k|+1,\zeta_k\rangle + g_k|-1,\zeta_k\rangle + h_k|(+1)P,\zeta_k\rangle + l_k|(-1)P,\zeta_k\rangle \right) \\
+ \sum_{k=1}^{N} \left(\bar{f}_k|1+,\bar{\zeta}_k\rangle + \bar{g}_k|1-,\bar{\zeta}_k\rangle + \bar{h}_k|(1+)P,\zeta_k\rangle + \bar{l}_k|(1-)P,\bar{\zeta}_k\rangle \right). \qquad (64)
\end{aligned}
$$

Using the values for the various inner products reported above, together with some simplifying algebra, we find the following expressions for the coefficients $f(\xi)$, etc.:

$$f(\xi) = -\frac{\langle 12,\xi|u\rangle}{S_{11}(\xi)^2}, \qquad (65)$$

$$g(\xi) = -\frac{\langle 13,\xi|u\rangle}{S_{11}(\xi)^2}, \qquad (66)$$

$$\bar{f}(\xi) = \frac{\langle 21,\xi|u\rangle}{\Delta_{11}(\xi)^2}, \qquad (67)$$

$$\bar{f}(\xi) = \frac{\langle 31,\xi|u\rangle}{\Delta_{11}(\xi)^2}, \qquad (68)$$

$$f_k = \frac{i}{S'_{11}(\zeta_k)^2}\langle (12)P,\zeta_k|u\rangle - \frac{2iS''_{11}(\zeta_k)}{S'_{11}(\zeta_k)^3}\langle 12,\zeta_k|u\rangle, \qquad (69)$$

$$g_k = \frac{i}{S'_{11}(\zeta_k)^2}\langle (13)P,\zeta_k|u\rangle - \frac{2iS''_{11}(\zeta_k)}{S'_{11}(\zeta_k)^3}\langle 13,\zeta_k|u\rangle, \qquad (70)$$

$$h_k = i\frac{\langle 12,\zeta_k|u\rangle}{S'_{11}(\zeta_k)^2}, \qquad (71)$$

$$l_k = i\frac{\langle 13,\zeta_k|u\rangle}{S'_{11}(\zeta_k)^2}, \qquad (72)$$

$$\bar{f}_k = \frac{i}{\Delta'_{11}(\zeta_k)^2}\langle (21)P,\bar{\zeta}_k|u\rangle - \frac{2i\Delta''_{11}(\zeta_k)}{\Delta'_{11}(\zeta_k)^3}\langle 21,\bar{\zeta}_k|u\rangle, \qquad (73)$$

$$\bar{g}_k = \frac{i}{\Delta'_{11}(\zeta_k)^2}\langle (31)P,\bar{\zeta}_k|u\rangle - \frac{2i\Delta''_{11}(\zeta_k)}{\Delta'_{11}(\zeta_k)^3}\langle 31,\bar{\zeta}_k|u\rangle, \qquad (74)$$

$$\bar{h}_k = i\frac{\langle 21,\bar{\zeta}_k|u\rangle}{\Delta'_{11}(\bar{\zeta}_k)^2}, \qquad (75)$$

$$\bar{l}_k = i\frac{\langle 31,\bar{\zeta}_k|u\rangle}{\Delta'_{11}(\bar{\zeta}_k)^2}. \qquad (76)$$

Here $S''_{11}(\zeta_k)$ means $d^2S_{11}/d\zeta^2|_{\zeta=\zeta_k}$, and similarly for $\Delta''_{11}(\bar{\zeta}_k)$.

Note that although we have expanded $|u\rangle$ in a basis containing $|\pm 1,\xi\rangle$, etc., the matrix elements which appear in the above expressions all contain an inner product of $|u\rangle$ with the "undressed" states $\langle 21,\xi|$, $\langle 31,\xi|$, etc.; in this form the expressions for these coefficients are most succinct. Since $|u\rangle$ is arbitrary, and if this basis is complete, we may use the above results to deduce an expression for the identity operator. This

is as follows:

$$
\begin{aligned}
\mathbf{I}_4 = &-\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{d\xi}{\Delta_{11}^2(\xi)}\left(|+1,\xi\rangle\langle 12,\xi| + |-1,\xi\rangle\langle 13,\xi|\right)\\
&+\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{d\xi}{S_{11}^2(\xi)}\left(|1+,\xi\rangle\langle 21,\xi| + |-1,\xi\rangle\langle 31,\xi|\right)\\
&+\sum_{k=1}^{N}\left(\frac{-4iS_{11}''(\zeta_k)}{S_{11}''(\zeta_k)^3}\,|+1,\zeta_k\rangle\langle 12,\zeta_k| + \frac{2i}{S_{11}'(\zeta_k)^2}|+1,\zeta_k\rangle\langle (12)P,\zeta_k|\right.\\
&\qquad -\frac{4iS_{11}''(\zeta_k)}{S_{11}'(\zeta_k)^3}|-1,\zeta_k\rangle\langle 13,\zeta_k| + \frac{2i}{S_{11}'(\zeta_k)^2}|-1,\zeta_k\rangle\langle (13)P,\zeta_k|\\
&\qquad \left.+\frac{2i}{S_{11}'(\zeta_k)^2}|(+1)P,\zeta_k\rangle\langle 12,\zeta_k| + \frac{2i}{S_{11}'(\zeta_k)^2}|(-1)P,\zeta_k\rangle\langle 13,\zeta_k|\right)\\
&+\sum_{k=1}^{N}\left(\frac{-4i\Delta_{11}''(\bar\zeta_k)}{\Delta_{11}''(\bar\zeta_k)^3}\,|1+,\bar\zeta_k\rangle\langle 21,\bar\zeta_k| + \frac{2i}{\Delta_{11}'(\bar\zeta_k)^2}|1+,\bar\zeta_k\rangle\langle (21)P,\bar\zeta_k|\right.\\
&\qquad -\frac{4i\Delta_{11}''(\bar\zeta_k)}{\Delta_{11}'(\bar\zeta_k)^3}|1-,\bar\zeta_k\rangle\langle 31,\bar\zeta_k| + \frac{2i}{\Delta_{11}'(\bar\zeta_k)^2}|1-,\bar\zeta_k\rangle\langle (31)P,\bar\zeta_k|\\
&\qquad \left.+\frac{2i}{\Delta_{11}'(\bar\zeta_k)^2}|(1+)P,\bar\zeta_k\rangle\langle 21,\bar\zeta_k| + \frac{2i}{\Delta_{11}'(\bar\zeta_k)^2}|(1-)P,\bar\zeta_k\rangle\langle 31,\bar\zeta_k|\right). \qquad (77)
\end{aligned}
$$

The above statement for the identity operator is a consequence of the development of our analysis this far, with a mixing of dressed and undressed states, and is not the final form for this operator.

The structure of (77) can be simplified on noting that quantities which appear in the first integral (i.e., S_{11}, $|+1,\xi\rangle$, etc.) are analytic in the upper half of the complex ζ plane, while all quantities in the second integral are analytic in the lower half plane. Then, (77) can be written in terms of two contour integrals,

$$
\begin{aligned}
\mathbf{I}_4 = &\frac{1}{\pi}\int_{\bar c}\left(|1+,\zeta\rangle\frac{1}{\Delta_{11}^2(\zeta)}\langle 21,\zeta| + |1-,\zeta\rangle\frac{1}{\Delta_{11}^2(\zeta)}\langle 31,\zeta|\right)d\zeta\\
&-\frac{1}{\pi}\int_{c}\left(|+1,\zeta\rangle\frac{1}{S_{11}^2(\zeta)}\langle 12,\zeta| + |-1,\zeta\rangle\frac{1}{S_{11}^2(\zeta)}\langle 13,\zeta|\right)d\zeta. \qquad (78)
\end{aligned}
$$

Here, $\bar c$ is the contour from $-\infty - i\epsilon$ to $+\infty - i\epsilon$ passing below all zeros of $\Delta_{11}(\zeta)$, whereas c is the contour from $-\infty + i\epsilon$ to $+\infty + i\epsilon$ passing above all zeros of $S_{11}(\zeta)$.

Equation (77) is recovered when the contours c and $\bar c$ are continued down to the real ζ axis, capturing contributions from the double poles from S_{11}^{-2} and Δ_{11}^{-2} in the process.

In the next section, we will prove that (78) is the correct identity operator for the space $L_2(-\infty,\infty)$. To simplify the structure of many of the resulting expressions, it remains expedient to continue with the use of contour integrals with contours c and $\bar c$ as defined above. However, we must first question whether or not it will always be possible to do this. In (85), for example, $S_{k1}(\zeta)$, $k = 2, 3$, only have the required analytic properties when the potentials q_1 and q_2 are on compact support. More generally, with $q_1,\ q_2 \in L_1$, $S_{k1}(\zeta)$ may have meaning only when ζ is real. The appropriate result for the latter case is obtained from (85) by first assuming that q_1 and q_2 are on compact support (so that S_{k1}, $k = 2, 3$, are analytic everywhere), then continuing the contour c to the real ζ axis, picking up residue contributions from the

poles of $S_{11}^{-1}(\zeta)$ in the process, and then relaxing the assumption of compact support. The final statement in the form of an integral along the real ζ axis together with a contribution from a discrete sum is the correct one for general q_1, q_2, ϵL_1.

All contour integrals over contours c and $\bar{c}$ written in the next section are to be interpreted in this manner.

4. Proof of completeness. To prove that (78) is indeed a statement of the identity operator, we will extend the approach used by Kaup in his discussion of the Zakharov–Shabat system, and use an appropriate form of a set of Marchenko equations for the problem to evaluate independently the right-hand side of (78).

Since the algebra can become quite horrendous—as witnessed by (64) to (78)—we will simplify the working by assuming that the potentials q_1 and q_2 are on compact support, but we stress that the final result is not dependent upon this assumption.

The basic idea is to construct a Riemann–Hilbert problem for the product states, which will be solved by recourse to an appropriate set of Marchenko equations. Results from this will then be used to evaluate the right-hand side of (78). The assumption of compact support permits a succinct formulation of the Riemann–Hilbert problem, enabling us to use it in (78). To illustrate this approach, we digress a little and first examine the corresponding Riemann–Hilbert problem for the simpler system described by (5). Each step here will correspond to a similar step in the more complicated problem, and will serve as a useful point of comparison.

The independent sets of Jost function solutions to (5) are related as stated in (8). With $i = 1$, this can be written as

$$\frac{\phi^{(1)}}{S_{11}} = \psi^{(1)} + \frac{S_{21}}{S_{11}}\psi^{(2)} + \frac{S_{31}}{S_{11}}\psi^{(3)}. \tag{79}$$

This is a Riemann–Hilbert problem, where the function on the left-hand side is analytic in the upper half plane $\Im\zeta \geq 0$, $\psi^{(1)}$ is analytic for $\Im\zeta < 0$, and the other two terms are analytic nowhere. To solve (79), we reformulate it as a Marchenko equation, following the standard procedure:

- Multiply (79) by $e^{i\xi t}/2\pi i(\xi - \bar{\zeta}_0)$, $\Im\bar{\zeta}_0 < 0$, and then integrate according to $\int_{-\infty}^{\infty} d\xi$, where $\xi = \Re\zeta$.
- Next, we follow references [8], [10] and argue that $\psi^{(i)}, i = 1, 2, 3$, can be written in the form

$$\psi^{(1)}(\zeta, t) = e^{-i\zeta t}\mathbf{e}_1 + \int_t^{\infty} e^{-i\zeta p}K^{(1)}(p, t)dp, \tag{80}$$

$$\psi^{(2)}(\zeta, t) = e^{i\zeta t}\mathbf{e}_2 + \int_t^{\infty} e^{i\zeta p}K^{(2)}(p, t)dp, \tag{81}$$

$$\psi^{(3)}(\zeta, t) = e^{i\zeta t}\mathbf{e}_3 + \int_t^{\infty} e^{i\zeta p}K^{(3)}(p, t)dp, \tag{82}$$

where the unit vectors $\mathbf{e}_i = (\delta_{i1}, \delta_{i2}, \delta_{i3})^T$, and the column vectors $K^{(i)}$ are independent of ζ. These are now substituted into (79) following application of the first step described above.

- Finally, we perform the integrals

$$\lim_{\epsilon\to 0}\int_{-\infty-i\epsilon}^{\infty-i\epsilon} e^{i\bar{\zeta}_0(\tau-t)}d\bar{\zeta}_0, \qquad \tau > t. \tag{83}$$

Then, (79) is cast into an equivalent equation relating the kernels $K^{(i)}, i = 1, 2, 3$,

$$K^{(1)}(\tau, t) = -F_2(t+\tau)\mathbf{e}_2 - F_3(t+\tau)\mathbf{e}_3 - \int_t^\infty F_2(p+\tau)K^{(2)}(p,t)dp - \int_t^\infty F_3(p+\tau)K^{(3)}(p,t)dp. \tag{84}$$

Here,

$$F_k(x) = \frac{1}{2\pi}\int_c \frac{S_{k1}(\zeta)}{S_{11}(\zeta)}e^{i\zeta x}d\zeta, \qquad k = 2, 3, \tag{85}$$

where the contour c is as defined in the previous section.

Alternatively, beginning with the equations

$$\frac{\phi^{(+)}}{\Delta_{11}} = \psi^{(2)} - \frac{\Delta_{21}}{\Delta_{11}}\psi^{(1)}, \tag{86}$$

$$\frac{\phi^{(-)}}{\Delta_{11}} = \psi^{(3)} - \frac{\Delta_{31}}{\Delta_{11}}\psi^{(1)}, \tag{87}$$

and following the same set of steps above, but now with $\bar{\zeta}_0$ replaced by ζ_0, where $\Im\zeta_0 > 0$, and with the limits of integration in the final step replaced by $-\infty + i\epsilon$ to $\infty + i\epsilon$, gives the required additional equations for $K^{(2)}$ and $K^{(3)}$,

$$K^{(2)}(\tau, t) = \bar{F}_2(\tau + t)\mathbf{e}_1 + \int_t^\infty \bar{F}_2(p+\tau)K^{(1)}(p,t)dp,$$

$$K^{(3)}(\tau, t) = \bar{F}_3(\tau + t)\mathbf{e}_1 + \int_t^\infty \bar{F}_3(p+\tau)K^{(1)}(p,t)dp, \tag{88}$$

where

$$\bar{F}_k = \frac{1}{2\pi}\int_{\bar{c}} \frac{\Delta_{k1}(\zeta)}{\Delta_{11}(\zeta)}e^{-i\zeta x}d\zeta, \qquad k = 2, 3. \tag{89}$$

Equations (84) and (88) are the Marchenko equations; one can show these have a unique solution using arguments similar to those used by Zakharov and Shabat elsewhere [7], [8]. The point we wish to emphasise here is that the existence of a unique solution to these equations is equivalent to the statement that (for $k = 2, 3$ and $\tau > t$)

$$\int_c \left(\psi^{(1)}(\zeta, t) + \frac{S_{21}(\zeta)}{S_{11}(\zeta)}\psi^{(2)}(\zeta, t) + \frac{S_{31}(\zeta)}{S_{11}(\zeta)}\psi^{(3)}(\zeta, t)\right)e^{i\zeta\tau}d\zeta = 0, \tag{90}$$

$$\int_{\bar{c}} \left(\psi^{(k)}(\zeta, t) - \frac{\Delta_{k1}(\zeta)}{\Delta_{11}(\zeta)}\psi^{(1)}(\zeta, t)\right)e^{-i\zeta\tau}d\zeta = 0 \tag{91}$$

or, more succinctly, that

$$\int_c \frac{1}{S_{11}(\zeta)}\phi^{(1)}(\zeta, t)e^{i\zeta\tau}d\zeta = 0, \qquad \tau > t, \tag{92}$$

$$\int_{\bar{c}} \frac{1}{\Delta_{11}(\zeta)}\phi^{(\pm)}(\zeta, t)e^{-i\zeta\tau}d\zeta = 0, \qquad \tau > t. \tag{93}$$

Now consider the product states, where the intention is again to link the existence and uniqueness of a set of Marchenko equations to a statement like (92) and (93) above. Using (8), (10) together with the definitions of $\hat{\phi}^{(\pm)}$ given in Appendix A, we have

$$\frac{\Gamma}{S_{11}^2}\left(\phi^{(1)}\wedge\hat{\phi}^{(+)}\right)^T = \psi^{(1)}\vee\hat{\psi}^{(2)} - \frac{S_{21}^2}{S_{11}^2}\psi^{(2)}\vee\hat{\psi}^{(1)} - \frac{S_{21}S_{31}}{S_{11}^2}\psi^{(3)}\vee\hat{\psi}^{(1)}$$

$$- \frac{S_{21}}{S_{11}}\psi^{(1)}\vee\hat{\psi}^{(1)} + \frac{S_{21}}{S_{11}}\psi^{(2)}\vee\hat{\psi}^{(2)} + \frac{S_{31}}{S_{11}}\psi^{(3)}\vee\hat{\psi}^{(2)}, \tag{94}$$

where the "metric" Γ—which converts vectors with $u\wedge\hat{v}$ structure into those with $u\vee\hat{v}$ structure—is the 4×4 matrix

$$\Gamma = \begin{pmatrix} 0 & \mathbf{I}_2 \\ -\mathbf{I}_2 & 0 \end{pmatrix}. \tag{95}$$

The term on the left-hand side of (94) is analytic in the half plane $\Im\zeta > 0$, while the first term on the right-hand side is analytic in the lower half plane; other terms (in the absence of compact support) are nowhere analytic. Equation (94) is a Riemann–Hilbert problem, and as such, will be studied following the outlined procedure. Step (i) requires multiplication of both sides by $e^{2i\xi t}/\left(\xi-\bar{\zeta}_0\right)$ for $\Im\bar{\zeta}_0 < 0$, then integration according to $(2\pi i)^{-1}\int_{-\infty}^{\infty} d\xi$, giving (after some rearrangement)

$$\begin{aligned}
&\psi^{(1)}\vee\hat{\psi}^{(2)}(\bar{\zeta}_0,t) = \mathbf{e}_1 e^{-2i\bar{\zeta}_0 t} \\
&-\frac{1}{2\pi i}\int_c\left(\frac{S_{21}^2}{S_{11}^2}\psi^{(2)}\vee\hat{\psi}^{(1)} + \frac{S_{21}S_{31}}{S_{11}^2}\psi^{(3)}\vee\hat{\psi}^{(1)}\right)\frac{e^{2i\left(\zeta-\bar{\zeta}_0\right)t}}{\zeta-\bar{\zeta}_0}d\zeta \\
&-\frac{1}{2\pi i}\int_c\left(\frac{S_{21}}{S_{11}}\psi^{(1)}\vee\hat{\psi}^{(1)} - \frac{S_{21}}{S_{11}}\psi^{(2)}\vee\hat{\psi}^{(2)} - \frac{S_{31}}{S_{11}}\psi^{(3)}\vee\hat{\psi}^{(2)}\right)\frac{e^{2i\left(\zeta-\bar{\zeta}_0\right)t}}{\zeta-\bar{\zeta}_0}d\zeta,
\end{aligned} \tag{96}$$

where $\mathbf{e}_i$ are now the four component unit vectors $\mathbf{e}_i = (\delta_{i1},\delta_{i2},\delta_{i3},\delta_{i4})^T$.

Here, of course, an assumption of compact support has been used to express (96) in the relatively compact form shown. Also, for clarity, all explicit reference to the ζ-dependence of the different terms in the integrals has been suppressed. The product states $\psi^{(1)}\vee\hat{\psi}^{(2)}$, $\psi^{(1)}\vee\hat{\psi}^{(3)}$,$\psi^{(2)}\vee\hat{\psi}^{(1)}$, and $\psi^{(3)}\vee\hat{\psi}^{(1)}$ will play the same role here as $\psi^{(1)}$, $\psi^{(2)}$, and $\psi^{(3)}$ played in the previous example. Consequently, we seek a set of coupled integral equations linking these vectors, which will then be turned into a set of Marchenko equations. First, however, we note the appearance of cross terms, such as $\psi^{(1)}\vee\hat{\psi}^{(1)}$, which need to be eliminated by relating them to the product eigenstates.

Consider the cross term $\psi^{(1)}\vee\hat{\psi}^{(1)}$. Substituting first for $\psi^{(1)}$ in terms of $\phi^{(i)}$, then for $\hat{\psi}^{(1)}$ in terms of $\hat{\phi}^{(i)}$, allows us to express $\psi^{(1)}\vee\hat{\psi}^{(1)}$ in two equivalent ways:

$$\psi^{(1)}\vee\hat{\psi}^{(1)} = \frac{\psi^{(1)}\vee\hat{\phi}^{(1)}}{\Delta_{11}} - \frac{\Delta_{21}}{\Delta_{11}}\psi^{(1)}\vee\hat{\psi}^{(2)} - \frac{\Delta_{31}}{\Delta_{11}}\psi^{(1)}\vee\hat{\psi}^{(3)} \tag{97}$$

$$= \frac{\phi^{(1)}\vee\hat{\psi}^{(1)}}{S_{11}} - \frac{S_{21}}{S_{11}}\psi^{(2)}\vee\hat{\psi}^{(1)} - \frac{S_{31}}{S_{11}}\psi^{(3)}\vee\hat{\psi}^{(1)}. \tag{98}$$

Note that the first term on each of the right sides is meromorphic in one or other half plane, with simple poles at the zeros of Δ_{11} or S_{11}. Note also that all remaining terms contain one or the other of the product eigenstates.

To proceed further, we now follow an argument used by Kaup; write

$$\psi^{(1)} \vee \hat{\psi}^{(1)}(\xi,t) = \frac{1}{2\pi i}\int_{\sigma} \frac{\psi^{(1)} \vee \hat{\psi}^{(1)}}{\zeta - \xi} d\zeta, \tag{99}$$

where σ denotes a small circular path centered at real ξ taken in the positive sense. Now substitute from (98) into the right-hand side of (99), using the first statement on that part of σ where $\Im\zeta > 0$, and the second where $\Im\zeta > 0$. Evaluating explicitly the integral of the first term in either case, and rearranging, then gives

$$\begin{aligned}\psi^{(1)} \vee \hat{\psi}^{(1)}(\zeta,t) &= \frac{1}{2\pi i}\int_{c}\left(\frac{S_{21}}{S_{11}}\psi^{(2)} \vee \hat{\psi}^{(1)} + \frac{S_{31}}{S_{11}}\psi^{(3)} \vee \hat{\psi}^{(1)}\right)\frac{d\zeta'}{\zeta' - \zeta} \\ &\quad - \frac{1}{2\pi i}\int_{\bar{c}}\left(\frac{\Delta_{21}}{\Delta_{11}}\psi^{(1)} \vee \hat{\psi}^{(2)} + \frac{\Delta_{31}}{\Delta_{11}}\psi^{(1)} \vee \hat{\psi}^{(3)}\right)\frac{d\zeta'}{\zeta' - \zeta}.\end{aligned} \tag{100}$$

A similar set of results relating the remaining cross product states in terms of the product eigenstates is quoted in Appendix B. Those can now be substituted into (96) giving our first integral statement linking the four product eigenstates. We do not show the explicit form for this first equation (i.e., that obtained by substituting for $\psi^{(1)} \vee \hat{\psi}^{(1)}$, $\psi^{(2)} \vee \hat{\psi}^{(2)}$, and $\psi^{(3)} \vee \hat{\psi}^{(2)}$) since it is rather cumbersome.

To complete the set of coupled integral equations for the product eigenstates, we require similar expressions for $\psi^{(1)} \vee \hat{\psi}^{(3)}$, $\psi^{(2)} \vee \hat{\psi}^{(1)}$, and $\psi^{(3)} \vee \hat{\psi}^{(1)}$, and these are

$$\begin{aligned}&\psi^{(1)} \vee \hat{\psi}^{(3)}(\bar{\zeta}_0,t) = \mathbf{e}_2 e^{-2i\bar{\zeta}_0 t} \\ &-\frac{1}{2\pi i}\int_{c}\left(\frac{S_{31}^2}{S_{11}^2}\psi^{(3)} \vee \hat{\psi}^{(1)} + \frac{S_{31}S_{21}}{S_{11}^2}\psi^{(2)} \vee \hat{\psi}^{(1)}\right)\frac{e^{2i(\zeta-\bar{\zeta}_0)t}}{\zeta - \bar{\zeta}_0}d\zeta \\ &-\frac{1}{2\pi i}\int_{c}\left(\frac{S_{31}}{S_{11}}\psi^{(1)} \vee \hat{\psi}^{(1)} - \frac{S_{21}}{S_{11}}\psi^{(2)} \vee \hat{\psi}^{(3)} - \frac{S_{31}}{S_{11}}\psi^{(3)} \vee \hat{\psi}^{(3)}\right)\frac{e^{2i(\zeta-\bar{\zeta}_0)t}}{\zeta - \bar{\zeta}_0}d\zeta,\end{aligned} \tag{101}$$

$$\begin{aligned}&\psi^{(2)} \vee \hat{\psi}^{(1)}(\zeta_0,t) = -\mathbf{e}_3 e^{2i\zeta_0 t} \\ &+\frac{1}{2\pi i}\int_{\bar{c}}\left(\frac{\Delta_{21}^2}{\Delta_{11}^2}\psi^{(1)} \vee \hat{\psi}^{(2)} + \frac{\Delta_{21}\Delta_{31}}{\Delta_{11}^2}\psi^{(1)} \vee \hat{\psi}^{(3)}\right)\frac{e^{-2i(\zeta-\zeta_0)t}}{\zeta - \zeta_0}d\zeta \\ &+\frac{1}{2\pi i}\int_{\bar{c}}\left(\frac{\Delta_{21}}{\Delta_{11}}\psi^{(1)} \vee \hat{\psi}^{(1)} - \frac{\Delta_{21}}{\Delta_{11}}\psi^{(2)} \vee \hat{\psi}^{(2)} - \frac{\Delta_{31}}{\Delta_{11}}\psi^{(2)} \vee \hat{\psi}^{(3)}\right)\frac{e^{-2i(\zeta-\zeta_0)t}}{\zeta - \zeta_0}d\zeta,\end{aligned} \tag{102}$$

$$\begin{aligned}&\psi^{(3)} \vee \hat{\psi}^{(1)}(\zeta_0,t) = -\mathbf{e}_4 e^{2i\zeta_0 t} \\ &+\frac{1}{2\pi i}\int_{\bar{c}}\left(\frac{\Delta_{31}^2}{\Delta_{11}^2}\psi^{(1)} \vee \hat{\psi}^{(3)} + \frac{\Delta_{21}\Delta_{31}}{\Delta_{11}^2}\psi^{(1)} \vee \hat{\psi}^{(2)}\right)\frac{e^{-2i(\zeta-\zeta_0)t}}{\zeta - \zeta_0}d\zeta \\ &+\frac{1}{2\pi i}\int_{\bar{c}}\left(\frac{\Delta_{31}}{\Delta_{11}}\psi^{(1)} \vee \hat{\psi}^{(1)} - \frac{\Delta_{31}}{\Delta_{11}}\psi^{(3)} \vee \hat{\psi}^{(3)} - \frac{\Delta_{21}}{\Delta_{11}}\psi^{(3)} \vee \hat{\psi}^{(2)}\right)\frac{e^{-2i(\zeta-\zeta_0)t}}{\zeta - \zeta_0}d\zeta.\end{aligned} \tag{103}$$

In each case, appropriate expressions are to be substituted for the cross states $\psi^{(1)} \vee \hat{\psi}^{(1)}$, $\psi^{(2)} \vee \hat{\psi}^{(2)}$, $\psi^{(3)} \vee \hat{\psi}^{(3)}$, $\psi^{(2)} \vee \hat{\psi}^{(3)}$, and $\psi^{(3)} \vee \hat{\psi}^{(2)}$ as required; the latter are listed in Appendix B.

These three equations, together with (96) then become four coupled integral equations relating the product eigenstates $\psi^{(2)} \vee \hat{\psi}^{(1)}$, etc.

To proceed further, we follow the argument that permitted the $\psi^{(i)}$ considered previously to be written in the form of (80)–(82). A similar reasoning shows that the product eigenstates can be written in the form

$$\psi^{(1)} \vee \hat{\psi}^{(2)}(\zeta,t) = \mathbf{e}_1 e^{-2i\zeta t} + \int_t^\infty e^{-2i\zeta p} M^{(1)}(t,p)dp, \tag{104}$$

$$\psi^{(1)} \vee \hat{\psi}^{(3)}(\zeta,t) = \mathbf{e}_2 e^{-2i\zeta t} + \int_t^\infty e^{-2i\zeta p} M^{(2)}(t,p)dp, \tag{105}$$

$$\psi^{(2)} \vee \hat{\psi}^{(1)}(\zeta,t) = -\mathbf{e}_3 e^{2i\zeta t} + \int_t^\infty e^{2i\zeta p} M^{(3)}(t,p)dp, \tag{106}$$

$$\psi^{(3)} \vee \hat{\psi}^{(1)}(\zeta,t) = -\mathbf{e}_4 e^{2i\zeta t} + \int_t^\infty e^{2i\zeta p} M^{(4)}(t,p)dp, \tag{107}$$

where the four-vectors $M^{(i)}$ exist, are unique, and are independent of the eigenparameter ζ.

These expressions are substituted into the integral equations (96), (101), and (103), as appropriate.

We now follow the final step of the procedure outlined at the beginning of this section and perform the integrals

$$\lim_{\epsilon\to 0} \int_{-\infty-i\epsilon}^{\infty-i\epsilon} e^{2i\bar{\zeta}_0(\tau-t)} d\bar{\zeta}_0, \qquad \tau > t, \tag{108}$$

on (96) and (101) and

$$\lim_{\epsilon\to 0} \int_{-\infty+i\epsilon}^{\infty+i\epsilon} e^{-2i\zeta_0(\tau-t)} d\zeta_0, \qquad \tau > t, \tag{109}$$

on (102) and (103), leading to the Marchenko equations

$$M^{(1)}(t,\tau) + \sum_{i=1}^4 F_i(t,\tau)\mathbf{e}_i + \sum_{i=1}^4 \int_t^\infty F_i(t,p)\Gamma_3 M^{(i)}(p,\tau)dp = 0, \tag{110}$$

$$M^{(2)}(t,\tau) + \sum_{i=1}^4 G_i(t,\tau)\mathbf{e}_i + \sum_{i=1}^4 \int_t^\infty G_i(t,p)\Gamma_3 M^{(i)}(p,\tau)dp = 0, \tag{111}$$

$$M^{(3)}(t,\tau) - \sum_{i=1}^4 \bar{F}_{i+2}(t,\tau)\mathbf{e}_i - \sum_{i=1}^4 \int_t^\infty \bar{F}_{i+2}(t,p)\Gamma_3 M^{(i)}(p,\tau)dp = 0, \tag{112}$$

$$M^{(4)}(t,\tau) - \sum_{i=1}^4 \bar{G}_{i+2}(t,\tau)\mathbf{e}_i - \sum_{i=1}^4 \int_t^\infty \bar{G}_{i+2}(t,p)\Gamma_3 M^{(i)}(p,\tau)dp = 0. \tag{113}$$

Here, Γ_3 is the 4×4 matrix

$$\Gamma_3 = \begin{pmatrix} \mathbf{I}_2 & 0 \\ 0 & -\mathbf{I}_2 \end{pmatrix} \tag{114}$$

and (scalar)

$$\begin{aligned} F_1(t,\tau) &= 2\nu_{22}(t,\tau) + \nu_{33}(t,\tau), \\ F_2(t,\tau) &= \nu_{23}(t,\tau), \\ F_3(t,\tau) &= 2\mu_{22}(t,\tau) + \gamma_2(t+\tau), \\ F_4(t,\tau) &= \mu_{23}(t,\tau) + \mu_{32}(t,\tau) + \lambda(t+\tau), \end{aligned} \tag{115}$$

$$
\begin{aligned}
G_1(t,\tau) &= \nu_{32}(t,\tau),\\
G_2(t,\tau) &= \nu_{22}(t,\tau) + 2\nu_{33}(t,\tau),\\
G_3(t,\tau) &= \mu_{23}(t,\tau) + \mu_{32}(t,\tau) + \lambda(t+\tau),\\
G_4(t,\tau) &= 2\mu_{33}(t,\tau) + \gamma_3(t+\tau).
\end{aligned} \tag{116}
$$

The scalar quantities $\bar{F}_i$ and $\bar{G}_i$ are defined similarly, but with all elements ν_{ij}, γ_i, etc., replaced with $\bar{\nu}_{ij}$, $\bar{\gamma}_i$, where these elements are defined as follows:

$$\mu_{ij}(x,y) = \frac{1}{2\pi^2 i}\int_c \frac{S_{i1}(\zeta)}{S_{11}(\zeta)} e^{2i\zeta x}\int_c \frac{S_{j1}(\zeta')}{S_{11}(\zeta')} e^{2i\zeta' y}\frac{d\zeta'}{\zeta'-\zeta-i\epsilon}d\zeta, \tag{117}$$

$$\nu_{ij}(x,y) = \frac{1}{2\pi^2 i}\int_c \frac{S_{i1}(\zeta)}{S_{11}(\zeta)} e^{2i\zeta x}\int_{\bar{c}} \frac{\Delta_{j1}(\zeta')}{\Delta_{11}(\zeta')} e^{-2i\zeta' y}\frac{d\zeta'}{\zeta'-\zeta}d\zeta, \tag{118}$$

$$\bar{\mu}_{ij}(x,y) = \frac{-1}{2\pi^2 i}\int_{\bar{c}} \frac{\Delta_{i1}(\zeta)}{\Delta_{11}(\zeta)} e^{-2i\zeta x}\int_{\bar{c}} \frac{\Delta_{j1}(\zeta')}{\Delta_{11}(\zeta')} e^{-2i\zeta' y}\frac{d\zeta'}{\zeta'-\zeta+i\epsilon}d\zeta, \tag{119}$$

$$\bar{\nu}_{ij}(x,y) = \frac{-1}{2\pi^2 i}\int_{\bar{c}} \frac{\Delta_{i1}(\zeta)}{\Delta_{11}(\zeta)} e^{-2i\zeta x}\int_c \frac{S_{j1}(\zeta')}{S_{11}(\zeta')} e^{2i\zeta' y}\frac{d\zeta'}{\zeta'-\zeta}d\zeta, \tag{120}$$

$$\gamma_j(x) = \frac{1}{\pi}\int_c \left(\frac{S_{j1}(\zeta)}{S_{11}(\zeta)}\right)^2 e^{2i\zeta x}d\zeta, \tag{121}$$

$$\bar{\gamma}_j(x) = \frac{1}{\pi}\int_{\bar{c}} \left(\frac{\Delta_{j1}(\zeta)}{\Delta_{11}(\zeta)}\right)^2 e^{-2i\zeta x}d\zeta, \tag{122}$$

$$\lambda(x) = \frac{1}{\pi}\int_c \frac{S_{21}(\zeta)S_{31}(\zeta)}{S_{11}^2(\zeta)} e^{2i\zeta x}d\zeta, \tag{123}$$

$$\bar{\lambda}(x) = \frac{1}{\pi}\int_{\bar{c}} \frac{\Delta_{21}(\zeta)\Delta_{31}(\zeta)}{\Delta_{11}^2(\zeta)} e^{-2i\zeta x}d\zeta. \tag{124}$$

The suffixes on $\bar{F}$ and $\bar{G}$ in (110)–(113) are defined modulo 4, so that $\bar{F}_5 \equiv \bar{F}_1$.

Equations (110)–(113) are the analogue of the Marchenko equations associated with the inverse to the scattering problem, (5), discussed previously. One can then argue (as before) that (110)–(113) have a unique solution. The point we now wish to emphasise is that the existence of a unique solution to this coupled system is equivalent to the statements that (cf. (92) and (93))

$$\int_c \frac{\phi^{(1)} \wedge \hat{\phi}^{(\pm)}(\zeta,t)}{S_{11}^2(\zeta)} e^{2i\zeta\tau}d\zeta = 0, \qquad \tau > t, \tag{125}$$

$$\int_{\bar{c}} \frac{\phi^{(\pm)} \wedge \hat{\phi}^{(1)}(\zeta,t)}{\Delta_{11}^2(\zeta)} e^{-2i\zeta\tau}d\zeta = 0, \qquad \tau > t. \tag{126}$$

These results will be used to evaluate the right-hand side of (78).

Return now to consider our statement of completeness, (78), which we write in the equivalent form

$$\mathbf{I}_4 = -\frac{1}{\pi}\int_c \left(\frac{\Delta_{22}}{S_{11}^3}|+1,\zeta\rangle\langle 1+,\zeta| + \frac{\Delta_{32}}{S_{11}^3}|+1,\zeta\rangle\langle 1-,\zeta| \right.$$
$$\left. + \frac{\Delta_{23}}{S_{11}^3}|-1,\zeta\rangle\langle 1+,\zeta| + \frac{\Delta_{33}}{S_{11}^3}|-1,\zeta\rangle\langle 1-,\zeta| \right) d\zeta$$
$$+\frac{1}{\pi}\int_{\bar{c}} \left(\frac{S_{22}}{\Delta_{11}^3}|1+,\zeta\rangle\langle +1,\zeta| + \frac{S_{32}}{\Delta_{11}^3}|1+\zeta\rangle\langle -1,\zeta| \right.$$
$$\left. + \frac{S_{23}}{\Delta_{11}^3}|1-,\zeta\rangle\langle +1,\zeta| + \frac{S_{33}}{\Delta_{11}^3}|1-\zeta\rangle\langle -1,\zeta| \right) d\zeta. \tag{127}$$

To simplify the notation, explicit reference to the ζ dependence of the parameters S_{ij} and Δ_{ij} has been suppressed.

Consider the matrix element $\langle\tau|\mathbf{I}|t\rangle$ of this equation. Using (35) and (36), the right-hand side of (127) becomes

$$\frac{1}{\pi}\int_c \frac{\Delta_{22}(\zeta)}{S_{11}^3(\zeta)}\psi^{(+)} \vee \hat{\psi}^{(1)}(\zeta,\tau)\left(\phi^{(1)} \wedge \hat{\phi}^{(+)}(\zeta,t)\right)^T d\zeta, \tag{128}$$

together with seven other terms corresponding to each of the entries in (127). We now substitute for $\phi^{(1)} \wedge \hat{\phi}^{(+)}$ from (94) and use (100), etc., to remove the cross terms. Finally, appealing to the Marchenko equations (110)–(113) and simplifying the ensuing expression gives

$$\langle\tau|\mathbf{I}|t\rangle = \mathbf{I}\delta\left(t-\tau\right). \tag{129}$$

Equivalently, in the last step of this proof we may appeal directly to (125) and (126) to deduce this result.

We have therefore shown that our statement (78) is indeed the required identity operator and that the basis discussed in connection with (64) is complete, provided that S_{11} and Δ_{11} have only simple zeros in their respective half planes. We now state without proof that (78) continues to be the identity operator for the more general case when the zeros of S_{11} and Δ_{11} are not necessarily simple. This may be shown to be true by retracing the derivation of the Marchenko equations for this latter case. If, for example, S_{11} has a double zero in the upper half ζ plane, the resolution of (78) into its continuum and discrete basis states would give additional basis elements corresponding to second and third derivatives of $|12,\zeta\rangle$, evaluated at the bound state eigenvalues. These extra contributions would then close the set.

Finally, we note that (127) applies for any potentials q_1, $q_2 \in L_1$; each integrand is meromorphic in the half space appropriate to the contours c and $\bar{c}$. Equation (127) can be expressed in terms of an integral along the real ζ axis together with a discrete sum of residues by continuing the contours c and $\bar{c}$ to the real ζ axis, as discussed at the end of section 3.

5. The potentials q_1 and q_2. We now use our completeness results to express the potentials q_1 and q_2 as appropriate linear combinations of the product states. Introduce $|Q\rangle$ such that

$$\langle t|Q\rangle = \begin{pmatrix} \mathbf{q} \\ \mathbf{q}^* \end{pmatrix}, \tag{130}$$

where $\mathbf{q} = (q_1, q_2)^T$. Using evolution equations for the components of the product states together with the known asymptotic forms of the product states as $t \to \pm\infty$,

we can show that

$$\langle 12,\zeta|Q\rangle = S_{11}(\zeta)\Delta_{12}(\zeta), \tag{131}$$

$$\langle 13,\zeta|Q\rangle = S_{11}(\zeta)\Delta_{13}(\zeta), \tag{132}$$

$$\langle 21,\zeta|Q\rangle = S_{12}(\zeta)\Delta_{11}(\zeta), \tag{133}$$

$$\langle 31,\zeta|Q\rangle = S_{13}(\zeta)\Delta_{11}(\zeta). \tag{134}$$

Then, using (78) to form $\langle t|\mathbf{I}_4|Q\rangle$ and expressing $\psi^{(\pm)}$ in terms of $\psi^{(1)}$ and $\psi^{(2)}$, we find

$$\begin{aligned}\begin{pmatrix} \mathbf{q} \\ \mathbf{q}^* \end{pmatrix} &= \frac{1}{\pi}\int_c \left(\frac{S_{21}}{S_{11}}\psi^{(2)} \vee \hat{\psi}^{(1)} + \frac{S_{31}}{S_{11}}\psi^{(3)} \vee \hat{\psi}^{(1)}\right) d\zeta \\ &\quad - \frac{1}{\pi}\int_{\bar{c}} \left(\frac{\Delta_{21}}{\Delta_{11}}\psi^{(1)} \vee \hat{\psi}^{(2)} + \frac{\Delta_{31}}{\Delta_{11}}\psi^{(1)} \vee \hat{\psi}^{(3)}\right) d\zeta.\end{aligned} \tag{135}$$

Writing these two integrals in terms of their discrete and continuum contributions gives the final expressions for q_1 and q_2 of $\mathbf{q}$:

$$\begin{aligned}\begin{pmatrix} q_1 \\ q_2 \end{pmatrix} &= \frac{1}{\pi}\int_{-\infty}^{\infty} \left(\frac{S_{21}}{S_{11}}\begin{pmatrix} \psi_1^{(2)}\hat{\psi}_2^{(1)} \\ \psi_1^{(2)}\hat{\psi}_3^{(1)} \end{pmatrix} + \frac{S_{31}}{S_{11}}\begin{pmatrix} \psi_1^{(3)}\hat{\psi}_2^{(1)} \\ \psi_1^{(3)}\hat{\psi}_3^{(1)} \end{pmatrix}\right) d\xi \\ &\quad - \frac{1}{\pi}\int_{-\infty}^{\infty} \left(\frac{\Delta_{21}}{\Delta_{11}}\begin{pmatrix} \psi_1^{(1)}\hat{\psi}_2^{(2)} \\ \psi_1^{(1)}\hat{\psi}_3^{(2)} \end{pmatrix} + \frac{\Delta_{31}}{\Delta_{11}}\begin{pmatrix} \psi_1^{(1)}\hat{\psi}_2^{(3)} \\ \psi_1^{(1)}\hat{\psi}_3^{(3)} \end{pmatrix}\right) d\xi \\ &\quad - 2i\sum_{k=1}^{N}\begin{pmatrix} \phi_1^{(1)}(\zeta_k)\hat{\psi}_2^{(1)}(\zeta_k) \\ \phi_1^{(1)}(\zeta_k)\hat{\psi}_3^{(1)}(\zeta_k) \end{pmatrix} \\ &\quad - 2i\sum_{k=1}^{\bar{N}=N}\begin{pmatrix} \psi_1^{(1)}(\bar{\zeta}_k)\hat{\phi}_2^{(1)}(\bar{\zeta}_k) \\ \psi_1^{(1)}(\bar{\zeta}_k)\hat{\phi}_3^{(1)}(\bar{\zeta}_k) \end{pmatrix},\end{aligned} \tag{136}$$

where the components $\phi_1^{(1)}(\zeta_k)$, etc., are the bound state Jost functions discussed previously.

Note that the system is not *diagonal* in the sense that both S_{21} and S_{31} (and Δ_{21} and Δ_{31}) contribute to both q_1 and q_2. However, in the linear limit, where both q_1 and q_2 are deemed to be small, we have $S_{11}(\xi) = \Delta_{11}(\xi) \sim 1$, while $\psi^{(2)} \sim \left(0, e^{i\xi t}, 0\right)^T$, etc., so that $\psi_1^{(1)}\hat{\psi}_2^{(2)} \sim e^{-2i\xi t}$ and $\psi_1^{(1)}\hat{\psi}_3^{(3)} \sim e^{-2i\xi t}$, with all other products negligible. Then (136) does uncouple, giving

$$q_1 = -\frac{1}{\pi}\int_{-\infty}^{\infty} e^{(-2i\xi t)}\Delta_{21}(\xi)d\xi, \tag{137}$$

$$q_2 = -\frac{1}{\pi}\int_{-\infty}^{\infty} e^{(-2i\xi t)}\Delta_{31}(\xi)d\xi, \tag{138}$$

as expected. Nonlinearity causes each of Δ_{21} and Δ_{31} to contribute to both q_1 and q_2. This will have interesting and tractable consequences for the (interesting) case when $\mathbf{q}$ corresponds to a vector soliton $\mathbf{q}_s$ accompanied by a small amount of radiation $\delta\mathbf{q}$; this is not pursued further here.

6. Final comments. We conclude this article with a few comments on the vector NLS equation (1). First we state the explicit forms for the conserved densities associated with (1). Though it is well known that such a set of conserved densities will be associated with this (integrable) equation, no explicit statement of this seems to have appeared in the literature. Second, we indicate how the matrix integro-differential operator (19) can be used to generate all other (monomial) members of the vector NLS equation. In particular, this should find useful application in studies on pulse propagation in birefringent fibers operated near the zero dispersion point, where the effects of third order dispersion become important. Kodama and Hasegawa [11] have shown that the scalar NLS equation with the effects of third order dispersion present at order $O(\epsilon)$ can be transformed into a higher order NLS equation in which a new perturbation on the integrable system now appears at order $O(\epsilon^2)$. It is anticipated that a similar transformation will exist for the vector system. Third, and finally, we show how the conserved densities, and the spectral data S_{ij}, evolve when an arbitrary perturbing term $i\mathbf{F}$ (which may correspond to weak and/or strong birefringence effects in the fiber, damping, third order dispersion, soliton self-frequency shift, etc.) is added to the right-hand side of (1).

The conserved densities are obtained by examining the asymptotics of $\phi_1^{(1)}$ as $t \to \infty$, together with the result that S_{11} does not vary with x. Denoting the conserved densities by C_n, $n = 0, 1, 2, \ldots$, it is thus found that

$$C_n = \int_{-\infty}^{\infty} \left(\mathbf{q}^T \rho_n\right) dt, \tag{139}$$

where

$$\rho_0 = \mathbf{q}^*,$$
$$\rho_1 = \mathbf{q}_t,$$

and

$$\rho_{n+1} = \rho_{n,t} + \sum_{i=0}^{n-1} \rho_i \mathbf{q}^T \rho_{n-i-1}. \tag{140}$$

Note that ρ_i are 2-component vectors.

In particular,

$$C_2 = \int_{-\infty}^{\infty} \left(\mathbf{q}^T \mathbf{q}_{tt}^* + \left(\mathbf{q}^\dagger \mathbf{q}\right)^2\right) dt \tag{141}$$

is the Hamiltonian functional for (1).

A general evolution equation for the vector NLS family can be expressed in the form

$$i\partial_x \begin{pmatrix} \mathbf{q} \\ -\mathbf{q}^* \end{pmatrix} - k(-2\mathcal{L}) \begin{pmatrix} \mathbf{q} \\ \mathbf{q}^* \end{pmatrix} = 0, \tag{142}$$

where $\mathcal{L}$ is the integro-differential operator from (19) and $k(\omega)$ is the dispersion function derived from the linearized problem with $\mathbf{q} \sim e^{i(\omega t - kx)}$.

For example, with

$$k(\omega) = -\omega^2 + \epsilon\omega^3$$

we get the integrable form of the vector NLS incorporating third order dispersion to be

$$i\mathbf{q}_x - \mathbf{q}_{tt} - 2\mathbf{q}^\dagger\mathbf{q}\mathbf{q} - i\epsilon\left(\mathbf{q}_{ttt} + 3\{\mathbf{q},\mathbf{q}^\dagger\}\mathbf{q}_t\right) = 0, \tag{143}$$

where $\{\mathbf{q},\mathbf{q}^\dagger\} = \mathbf{q}\mathbf{q}^\dagger + \mathbf{q}^\dagger\mathbf{q}$ denotes the anticommutator.

Finally, we note that for the perturbed form of the vector NLS equation

$$i\mathbf{q}_x - \mathbf{q}_{tt} - 2\mathbf{q}^\dagger\mathbf{q}\mathbf{q} = i\mathbf{F} \tag{144}$$

with $\mathbf{F} = (F_1, F_2)^T$, the conserved quantities C_n evolve according to the equations (cf. [12] for the scalar case)

$$\frac{d}{dx}C_n = \int_{-\infty}^{\infty}\left(\mathbf{F}^T, -\mathbf{F}\right)\left(\mathcal{L}_A\right)^n\begin{pmatrix}\mathbf{q}^*\\ -\mathbf{q}\end{pmatrix}dt, \tag{145}$$

where $\mathcal{L}_A$ is the adjoint integro-differential operator from (29). Similarly, the spectral data now evolve in accordance with [13],

$$S_{ij,x} = S^0_{ij,x} + \int_{-\infty}^{\infty}\phi^{(j)}\wedge\hat{\psi}^{(i)}\begin{pmatrix}\mathbf{F}\\ -\mathbf{F}^*\end{pmatrix}dt, \tag{146}$$

where the term $S^0_{ij,x}$ represents the usual evolution for the unperturbed system. In particular, $S^0_{i1,x} = -4i\zeta^2 S^0_{i1}$, $i = 2, 3$, while $S^0_{11,x} = 0$.

Specific forms for the perturbation include the following:

- $\beta\sigma_3\mathbf{q}$, where β is a scalar and σ_3 is a Pauli matrix; cf. (28). This term models the weak birefringence properties of the fiber and results in a difference in phase velocity between pulse components in the two polarization eigenmodes by an amount 2β.
- $i\beta'\sigma_3\mathbf{q}_t$, where β' is a constant. This term models the strong birefringence properties of the fiber, where now the group velocities of the two modes differ by $\pm\beta'$.
- $\Gamma\mathbf{q}$, where Γ is a (real) constant. This term models loss on the fiber.
- $i\epsilon\mathbf{q}_{ttt}$. This corresponds to higher order dispersive effects in the fiber.

Other terms corresponding to different interactive, optical processes could similarly be quoted.

Suppose now that a soliton is inserted into the fiber at $x = 0$. Then, $S_{ij}(\zeta, x = 0) = 0$ for all $i \neq j$. Under the action of any of the above perturbing influences, radiation will be shed by the soliton on propagating down the fiber so that the pulse acquires a structure $\mathbf{q}(x,t) = \mathbf{q}_s(x,t) + \delta\mathbf{q}(x,t)$, say, where the suffix s denotes the soliton contribution.

Using (146), $S_{ij}(\zeta, x)$ can be computed for some specific form for $\mathbf{F}$, assuming that $\phi^{(j)}\wedge\hat{\psi}^i$ is known. If the perturbing influence is weak, the product states can be approximated by their soliton expressions; the integral on the right-hand side of (146) is then evaluated and $S_{ij}(\zeta, x)$ obtained.

On continuing the contours c and $\bar{c}$ to the real axis, the continuum contribution is identified as yielding $\delta\mathbf{q}(x,t)$, so that

$$\begin{aligned}\begin{pmatrix}\delta q_1\\ \delta q_2\end{pmatrix} &= \frac{1}{\pi}\int_{-\infty}^{\infty}\left(\frac{S_{21}}{S_{11}}\begin{pmatrix}\psi_1^{(2)}\hat{\psi}_2^{(1)}\\ \psi_1^{(2)}\hat{\psi}_3^{(1)}\end{pmatrix} + \frac{S_{31}}{S_{11}}\begin{pmatrix}\psi_1^{(3)}\hat{\psi}_2^{(1)}\\ \psi_1^{(3)}\hat{\psi}_3^{(1)}\end{pmatrix}\right)d\xi\\ &\quad - \frac{1}{\pi}\int_{-\infty}^{\infty}\left(\frac{S^*_{21}}{S^*_{11}}\begin{pmatrix}\psi_1^{(1)}\hat{\psi}_2^{(2)}\\ \psi_1^{(1)}\hat{\psi}_3^{(2)}\end{pmatrix} + \frac{S^*_{31}}{S^*_{11}}\begin{pmatrix}\psi_1^{(1)}\hat{\psi}_2^{(3)}\\ \psi_1^{(1)}\hat{\psi}_3^{(3)}\end{pmatrix}\right)d\xi.\end{aligned} \tag{147}$$

Inserting the derived expressions for S_{21}, etc., and approximating the components $\psi_j^{(i)}$, $\hat{\psi}_j^i$ in these integrals with their solitonic expressions as before then leads to the required expressions for δq_1 and δq_2—at least in principle.

In fiber optical communication systems utilizing solitons, it is important to know how radiation is shed by a soliton in a birefringent fiber and how this radiation propagates away from the main pulse, since it can then interact with other solitons in the fiber. The mathematical formalism developed here resulting in (147) should find useful application to that end.

Appendix A. The dressed Jost functions $\phi^{(\pm)}$etc. In each pair of equations below, the first equation defines the corresponding quantity while the second follows from substitution from (8) and (10) as appropriate.

We define

$$\begin{aligned}
\phi^{(+)} &= S_{33}\phi^{(2)} - S_{32}\phi^{(3)} \\
&= \Delta_{11}\psi^{(2)} - \Delta_{21}\psi^{(1)}, && \text{(A1)} \\
\phi^{(-)} &= S_{22}\phi^{(3)} - S_{23}\phi^{(2)} \\
&= \Delta_{11}\psi^{(2)} - \Delta_{31}\psi^{(1)}, && \text{(A2)} \\
\hat{\phi}^{(+)} &= \Delta_{33}\hat{\phi}^{(2)} - \Delta_{32}\hat{\phi}^{(3)} \\
&= S_{11}\hat{\psi}^{(2)} - S_{21}\hat{\psi}^{(1)}, && \text{(A3)} \\
\hat{\phi}^{(-)} &= \Delta_{22}\hat{\phi}^{(3)} - \Delta_{23}\hat{\phi}^{(2)} \\
&= S_{11}\hat{\psi}^{(3)} - S_{31}\hat{\psi}^{(1)}, && \text{(A4)} \\
\psi^{(+)} &= \Delta_{33}\psi^{(2)} - \Delta_{23}\psi^{(3)} \\
&= S_{11}\phi^{(2)} - S_{12}\phi^{(1)}, && \text{(A5)} \\
\psi^{(-)} &= \Delta_{22}\psi^{(3)} - \Delta_{32}\psi^{(2)} \\
&= S_{11}\phi^{(3)} - S_{13}\phi^{(1)}, && \text{(A6)} \\
\hat{\psi}^{(+)} &= S_{33}\hat{\psi}^{(2)} - S_{23}\hat{\psi}^{(3)} \\
&= \Delta_{11}\hat{\phi}^{(2)} - \Delta_{12}\hat{\phi}^{(1)}, && \text{(A7)} \\
\hat{\psi}^{(-)} &= S_{22}\hat{\psi}^{(3)} - S_{32}\hat{\psi}^{(2)} \\
&= \Delta_{11}\hat{\phi}^{(3)} - \Delta_{13}\hat{\phi}^{(1)}. && \text{(A8)}
\end{aligned}$$

Appendix B. Expressions for the product cross states in terms of the product eigenstates. For each case, we follow the working for $\psi^{(1)} \vee \hat{\psi}^{(1)}$ discussed in the text.

(i) $\psi^{(2)} \vee \hat{\psi}^{(2)}$. Substituting in turn for $\psi^{(2)}$ and $\hat{\psi}^{(2)}$ by using expressions given in Appendix A allows us to write $\psi^{(2)} \vee \hat{\psi}^{(2)}$ in two equivalent ways:

$$\begin{aligned}
\psi^{(2)} \vee \hat{\psi}^{(2)} &= \frac{\phi^{(+)} \vee \hat{\psi}^{(2)}}{\Delta_{11}} - \frac{\Delta_{21}}{\Delta_{11}}\psi^{(1)} \vee \hat{\psi}^{(2)} \\
&= \frac{\psi^{(2)} \vee \hat{\phi}^{(+)}}{S_{11}} - \frac{S_{21}}{S_{11}}\psi^{(2)} \vee \hat{\psi}^{(1)}. && \text{(B1)}
\end{aligned}$$

Here again, we observe that the first terms on the right-hand sides are meromorphic in one or the other half place, with simple poles at the zeros of Δ_{11} or S_{11}. Note also that the remaining terms contain one or other of the product eigenstates.

Following (99), we write

$$\psi^{(2)} \vee \hat{\psi}^{(2)}(\xi) = \frac{1}{2\pi i} \int_{\sigma} \frac{\psi^{(2)} \vee \hat{\psi}^{(2)}}{\zeta - \xi} d\zeta, \tag{B2}$$

where σ denotes a small circular path centered at (real) ξ, taken in the positive sense. Now substitute (B1) into (B2), using the first statement in (B1) on the semicircular path with $\Im\zeta < 0$ and the second on $\Im\zeta > 0$. Evaluating explicitly the first integral in either case and rearranging give

$$\psi^{(2)} \vee \hat{\psi}^{(2)}(\zeta) = \frac{1}{2\pi i} \int_{c} \frac{S_{21}}{S_{11}} \psi^{(2)} \vee \hat{\psi}^{(1)} \frac{d\zeta'}{\zeta' - \zeta} - \frac{1}{2\pi i} \int_{\bar{c}} \frac{\Delta_{21}}{\Delta_{11}} \psi^{(1)} \vee \hat{\psi}^{(2)} \frac{d\zeta'}{\zeta' - \zeta}. \tag{B3}$$

Similarly, for each of the other cross product states, the corresponding statements read as follows:

(ii) $\psi^{(3)} \vee \hat{\psi}^{(2)}$:

$$\begin{aligned} \psi^{(3)} \vee \hat{\psi}^{(2)}(\zeta) &= \frac{\psi^{(3)} \vee \hat{\phi}^{(+)}}{S_{11}} - \frac{S_{21}}{S_{11}} \psi^{(3)} \vee \hat{\psi}^{(1)} \\ &= \frac{\phi^{(-)} \vee \hat{\psi}^{(2)}}{\Delta_{11}} - \frac{\Delta_{31}}{\Delta_{11}} \psi^{(1)} \vee \hat{\psi}^{(2)} \end{aligned} \tag{B4}$$

leading to

$$\psi^{(3)} \vee \hat{\psi}^{(2)}(\zeta) = \frac{1}{2\pi i} \int_{c} \frac{S_{21}}{S_{11}} \psi^{(3)} \vee \hat{\psi}^{(1)} \frac{d\zeta'}{\zeta' - \zeta} - \frac{1}{2\pi i} \int_{\bar{c}} \frac{\Delta_{31}}{\Delta_{11}} \psi^{(1)} \vee \hat{\psi}^{(2)} \frac{d\zeta'}{\zeta' - \zeta}; \tag{B5}$$

(iii) $\psi^{(2)} \vee \hat{\psi}^{(3)}$:

$$\begin{aligned} \psi^{(2)} \vee \hat{\psi}^{(3)}(\zeta) &= \frac{\phi^{(+)} \vee \hat{\psi}^{(3)}}{\Delta_{11}} - \frac{\Delta_{21}}{\Delta_{11}} \psi^{(1)} \vee \hat{\psi}^{(3)} \\ &= \frac{\psi^{(2)} \vee \hat{\phi}^{(-)}}{S_{11}} - \frac{S_{31}}{S_{11}} \psi^{(2)} \vee \hat{\psi}^{(1)} \end{aligned} \tag{B6}$$

leading to

$$\psi^{(2)} \vee \hat{\psi}^{(3)}(\zeta) = \frac{1}{2\pi i} \int_{c} \frac{S_{31}}{S_{11}} \psi^{(2)} \vee \hat{\psi}^{(1)} \frac{d\zeta'}{\zeta' - \zeta} - \frac{1}{2\pi i} \int_{\bar{c}} \frac{\Delta_{21}}{\Delta_{11}} \psi^{(1)} \vee \hat{\psi}^{(3)} \frac{d\zeta'}{\zeta' - \zeta}; \tag{B7}$$

(iv) $\psi^{(3)} \vee \hat{\psi}^{(3)}$:

$$\begin{aligned} \psi^{(3)} \vee \hat{\psi}^{(3)}(\zeta) &= \frac{\psi^{(3)} \vee \hat{\phi}^{(-)}}{S_{11}} - \frac{S_{31}}{S_{11}} \psi^{(3)} \vee \hat{\psi}^{(1)} \\ &= \frac{\phi^{(-)} \vee \hat{\psi}^{(3)}}{\Delta_{11}} - \frac{\Delta_{31}}{\Delta_{11}} \psi^{(1)} \vee \hat{\psi}^{(3)} \end{aligned} \tag{B8}$$

leading to

$$\psi^{(3)} \vee \hat{\psi}^{(3)}(\zeta) = \frac{1}{2\pi i} \int_{c} \frac{S_{31}}{S_{11}} \psi^{(3)} \vee \hat{\psi}^{(1)} \frac{d\zeta'}{\zeta' - \zeta} - \frac{1}{2\pi i} \int_{\bar{c}} \frac{\Delta_{31}}{\Delta_{11}} \psi^{(1)} \vee \hat{\psi}^{(3)} \frac{d\zeta'}{\zeta' - \zeta}. \tag{B9}$$

In each of the above integrals, the explicit ζ' dependence of each term in the integrand has been suppressed for clarity.

When the expressions for the different cross product states (together with (100) for $\psi^{(1)} \vee \hat{\psi}^{(1)}$) are substituted into (96) and (101)–(103), the resulting equations will be a set of four integral equations coupling the product eigenstates $\psi^{(1)} \vee \hat{\psi}^{(2)}$, etc.

Acknowledgment. One of the authors (R.G.D.) is pleased to acknowledge financial support in the form of a postgraduate studentship from the EPSRC.

REFERENCES

[1] S.V. Manakov, *On the theory of two-dimensional stationary self-focusing of electromagnetic waves*, Soviet Phys. JETP, 8 (1974), pp. 248–253.
[2] A. Hasegawa and Y. Kodama, *Optical Communications*, Clarendon Press, Oxford, UK, 1995.
[3] C.R. Menyuk, *Stability of solitons in birefringent optical fibers.* I. *Equal propagation amplitudes*, Opt. Lett., 12 (1987), pp. 614–616.
[4] C.R. Menyuk, *Stability of solitons in birefringent optical fibers.* I. *Equal propagation amplitudes*, J. Opt. Soc. Amer. B, 5 (1988), pp. 392–402.
[5] D. Marcuse, C.R. Menyuk, and P.K.A. Wai, *Application of the Manakov-PMD equations to studies of signal propagation in fibers with randomly varying birefringence*, J. Lightwave Technology, 15 (1997), pp. 1755–1745.
[6] P.K.A. Wai and C.R. Menyuk, *Polarization mode dispersion, decorrelation and diffusion in optical fibers with randomly varying birefringence*, J. Lightwave Technology, 14 (1996), pp. 148–157.
[7] M.J. Ablowitz and H. Segur, *Solitons and the Inverse Scattering Transform*, SIAM Stud. Appl. Math. 4, SIAM, Philadelphia, 1981.
[8] V.E. Zakharov and P.B. Shabat, *Exact theory of two-dimensional self-focusing and one-dimensional self-modulation of waves in nonlinear media*, Soviet Phys. JETP, 34 (1972), pp. 62–69.
[9] D.J. Kaup, *Closure of the squared Zakharov-Shabat eigenstates*, J. Math. Anal. Appl., 54 (1976), pp. 849–864.
[10] A.I. Maimistov, A.M. Basharov, S.O. Elyutin, and Yu. M. Sklyarov, *Present state of self-induced transparency theory*, Phys. Rep., 191 (1990), pp. 1–108.
[11] Y. Kodama and A. Hasegawa, *Nonlinear pulse propagation in a monomode dielectric guide*, IEEE J. Quantum Electron., 23 (1987), p. 510.
[12] J.N. Elgin, *Perturbations of optical solitons*, Phys. Rev. A, 47 (1993), pp. 4331–4341.
[13] J.N. Elgin, *Perturbed solitons in birefringent fibers*, in Physics and Applications of Optical Solitons in Fibers, A. Hasegawa, ed., Kluwer Academic Publishing, Norwell, MA, 1996, pp. 53–58.

SIAM J. MATH. ANAL.
Vol. 32, No. 1, pp. 80–102

ANGULAR CHANNELS IN A MULTIDIMENSIONAL WAVELET TRANSFORM*

ERIC CLARKSON†

Abstract. Given a subgroup S of $GL(n)$, let G be the semidirect product of S with $\mathbb{R}^n$. The wavelet transform is defined for functions in $L^2(\mathbb{R}^n)$ by using the action of G on this space. The standard properties of the wavelet transform and its inverse are quickly and easily derived in this formalism. In particular, the admissibility condition for the wavelet is expressed in terms of an integral over S. The notion of orthogonal wavelet channels is defined, and the wavelet transform is decomposed in terms of them. Other operators on $L^2(\mathbb{R}^n)$ can also be analyzed in terms of their mixing of wavelet channels. For $n = 2$ and $n = 3$, details are given for the expansion of an arbitrary wavelet transform in terms of angular wavelet channels. An example is provided for $n = 2$. The correspondence between angular channels and the spherical harmonic decomposition of the Fourier transform of the wavelet transform is also outlined.

Key words. wavelet transform, unitary representations of locally compact groups, orthogonal functions, spherical harmonics, signal reconstruction, channel models

AMS subject classifications. 22D10, 43A15, 44A05, 94A11, 94A12, 94A40

PII. S0036141096309617

1. Introduction.

Wavelet transforms for functions of more than one variable have been approached from various directions. One method has been to use the theory of square integrable representations for nonunimodular, locally compact groups (see [4, 5, 7, 12]), which is very general, and apply it to a given symmetry group acting on $\mathbb{R}^n$. This has been done for the euclidean symmetry group in [2, 11]. Another method is to restrict the point symmetry group to be n-dimensional and use elementary arguments, as in [3]. This is good for $n = 2$, but for $n = 3$ the rotation-dilation group, an important example, is four-dimensional. Elementary arguments are also possible if the group contains only scale and shift operations. Optical realizations of this type of transform for $n = 2$ are carried out in [1, 10, 14, 15, 16]. In image analysis applications, however, orientation is a useful parameter. The wavelet transform derivations given below fall in the midrange. Since integration on the group is used, the arguments are not exactly elementary. On the other hand, the symmetry groups considered are all semidirect products of a translation groups $T = \mathbb{R}^n$ with a subgroup S of $GL(n)$. This is a small subclass of the set of all locally compact, nonunimodular groups, but it does cover all of the special cases mentioned above.

The decomposing of a wavelet transform into orthogonal channels in the second half of this article seems to be new, although there are hints of it in [9, 13]. For an actual image analysis system that uses the wavelet transform, it would be advantageous to have a variety of wavelets available. By using orthogonal wavelet channels, an infinite collection of wavelet transforms could be synthesized by varying some parameters, or one wavelet transform can be computed and then others derived from it by projection. In the special case of angular wavelet channels, the parameters can be

*Received by the editors September 20, 1996; accepted for publication September 9, 1998; published electronically June 22, 2000. This publication was made possible by National Cancer Institute grant R01-CA52643. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the National Cancer Institute.

http://www.siam.org/journals/sima/32-1/30961.html

†University of Arizona Health Sciences Center, Department of Radiology Research, Tucson, AZ 85724 (clarksoneric@netscape.net).

varied to emphasize certain angular features of the object. If reconstruction is also necessary, angular channels can be used to eliminate the integration over the rotation group that is normally necessary when rotations are included in the symmetry operations. Sampling can also be simplified by using a finite number of angular channels of a particular wavelet and sampling each of them in scale and translation only. Finally, in the Fourier domain, angular channels correspond to an expansion of the Fourier transform of the wavelet transform in spherical harmonics in the angular variables.

2. The one-dimensional wavelet transform. As is well known the one-dimensional wavelet transform may be formulated in terms of the one-dimensional affine group (see [6, 8]). This is reviewed in this section in order to establish notation and provide a pattern for generalizing to higher dimensions. In particular, the reformulation of the admissibility condition in terms of the scale group, while trivial in one dimension, leads to a convenient condition for admissibility in higher dimensions.

The wavelet transform of $f \in L^2(\mathbb{R})$ with respect to $h \in L^2(\mathbb{R})$ is given by

$$w(s,t) = \int_{\mathbb{R}} f(x) \frac{1}{\sqrt{|s|}} h^* \left(\frac{x-t}{s} \right) dx.$$

If h is admissible, f may be recovered via

$$f(x) = \frac{1}{C_h} \int_{\mathbb{R}^2} w(s,t) \frac{1}{\sqrt{|s|}} h \left(\frac{x-t}{s} \right) \frac{1}{s^2} ds dt.$$

The admissibility condition is

$$C_h = \int_{\mathbb{R}} |H(k)|^2 \frac{1}{|k|} dk < \infty.$$

$H(k)$ is the Fourier transform of $h(x)$. This convention will be used for other functions also. For our purposes, this can be rewritten trivially as

$$C_h = \int_{\mathbb{R}} |H(sk)|^2 \frac{1}{|s|} ds < \infty.$$

Let S be the multiplicative group of nonzero real numbers and T the additive group of real numbers. The one-dimensional affine group G is the semidirect product of S with T. An element of G is $g = (s,t)$, with $s \in S$ and $t \in T$, and if $g' = (s',t')$, then $gg' = (ss', st' + t)$. The standard action of G on $\mathbb{R}$ is given by

$$\Phi_g(x) = sx + t,$$

giving $\Phi_g \Phi_{g'} = \Phi_{gg'}$ and $\Phi_g^{-1}(x) = \Phi_{g^{-1}}(x) = \frac{x-t}{s}$. Using this action, elements of G may also act on functions such as $h(x)$ above:

$$\overline{\Phi}_g h(x) = \frac{1}{\sqrt{|s|}} h\left(\Phi_g^{-1}(x)\right) = \frac{1}{\sqrt{|s|}} h\left(\frac{x-t}{s} \right).$$

Again $\overline{\Phi}_g \overline{\Phi}_{g'} = \overline{\Phi}_{gg'}$, making $\overline{\Phi}$ an irreducible unitary representation of G on $L^2(\mathbb{R})$.

The left invariant measure (left Haar measure) on G will be denoted by dg_L. If w is a function on G, integrable with respect to this measure, and $g_0 \in G$, then the defining property of the left invariant measure for any group is

$$\int_G w(g_0 g) dg_L = \int_G w(g) dg_L.$$

If G is the one-dimensional affine group, then

$$\frac{dsdt}{s^2} = dg_L.$$

If ds_L is the left invariant measure on S, then

$$\frac{ds}{|s|} = ds_L.$$

Now the wavelet transform formulae may be written

$$w(g) = \int_{\mathbb{R}} f(x)\overline{\Phi}_g h^*(x)dx, \tag{1}$$

$$f(x) = \frac{1}{C_h}\int_G w(g)\overline{\Phi}_g h(x)dg_L, \tag{2}$$

$$C_h = \int_S |H(sk)|^2 \, ds_L < \infty. \tag{3}$$

3. A multidimensional wavelet transform. These last three formulae are easily scaled up to higher dimensions. However, there are now two questions that must be addressed. What group should be used for S? Is C_h independent of $\mathbf{k}$ for a particular choice for S? The answer to the second question is yes if the chosen group S satisfies an orbit condition in $\mathbb{R}^n$. The answer to the first question depends upon the application.

T is now the n-dimensional translation group (isomorphic to $\mathbb{R}^n$) and S is a subgroup of $GL(n)$. As before $G = \{(s, \mathbf{t}) : s \in S, \mathbf{t} \in T\}$ with multiplication rule $(s, \mathbf{t})(s', \mathbf{t}') = (ss', s\mathbf{t} + \mathbf{t}')$. Let $J(s) = |\det(s)|$. Again for each $g \in G$ we have operators on $\mathbb{R}^n$ and $L^2(\mathbb{R}^n)$

$$\Phi_g \mathbf{x} = s\mathbf{x} + \mathbf{t},$$

$$\overline{\Phi}_g h(\mathbf{x}) = \frac{1}{\sqrt{J(s)}} h\left(\Phi_g^{-1}\mathbf{x}\right) = \frac{1}{\sqrt{J(s)}} h\left(s^{-1}(\mathbf{x} - \mathbf{t})\right).$$

PROPOSITION 1. *Given $f, h \in L^2(\mathbb{R}^n)$ let*

$$w(g) = \int_{\mathbb{R}^n} f(\mathbf{x})\overline{\Phi}_g h^*(\mathbf{x})d^n x \tag{4}$$

and, for $\mathbf{k} \neq 0$,

$$C_h(\mathbf{k}) = \int_S \left|H\left(s^t\mathbf{k}\right)\right|^2 ds_L. \tag{5}$$

If $C_h(\mathbf{k})$ is finite and independent of $\mathbf{k}$, then

$$f(\mathbf{x}) = \frac{1}{C_h}\int_G w(g)\overline{\Phi}_g h(\mathbf{x})dg_L. \tag{6}$$

Proof. Let $g = (s, \mathbf{t})$ as above. Then

$$dg_L = \left(\frac{d^n t}{J(s)} \right) ds_L,$$

and the forward transform is

$$w(s, \mathbf{t}) = \int_{\mathbb{R}^n} f(\mathbf{x}) \frac{1}{\sqrt{J(s)}} h^* \left(s^{-1}(\mathbf{x} - \mathbf{t}) \right) d^n x. \tag{7}$$

As will be seen below, this wavelet transform maps $L^2(\mathbb{R}^n)$ to $L^2(G)$, the Hilbert space of functions on G square integrable with respect to dg_L. Let $\widetilde{f}$ be given by the adjoint of this transform applied to w:

$$\begin{aligned} \widetilde{f}(\mathbf{x}) &= \int_G w(g) \overline{\Phi}_g h(\mathbf{x}) dg_L \\ &= \int_S \int_{\mathbb{R}^n} w(s, \mathbf{t}) \frac{1}{\sqrt{J(s)}} h\left(s^{-1}(\mathbf{x} - \mathbf{t}) \right) \frac{d^n t}{J(s)} ds_L, \end{aligned} \tag{8}$$

and define $h_s(\mathbf{x}) = \frac{1}{\sqrt{J(s)}} h\left(s^{-1}\mathbf{x} \right)$ and $w_s(\mathbf{t}) = w(s, \mathbf{t})$. Taking the Fourier transform of (7) with respect to $\mathbf{t}$ gives

$$W_s(\mathbf{k}) = F(\mathbf{k}) H_s^*(\mathbf{k}) = F(\mathbf{k}) H^*(s^t \mathbf{k}) \sqrt{J(s)}.$$

Now let $\mathcal{F}$ denote the Fourier transform operator on $L^2(\mathbb{R}^n)$. Then (8) may be rewritten as

$$\widetilde{f}(\mathbf{x}) = \int_S \mathcal{F}^{-1} \{ W_s H_s \} \frac{ds_L}{J(s)} = \int_S \mathcal{F}^{-1} \{ F H_s^* H_s \} \frac{ds_L}{J(s)},$$

which gives

$$\begin{aligned} \widetilde{f}(\mathbf{x}) &= \int_S \int_{\mathbb{R}^n} F(\mathbf{k}) H^* \left(s^t \mathbf{k} \right) H \left(s^t \mathbf{k} \right) e^{2\pi i \mathbf{k} \cdot \mathbf{x}} d^n k ds_L \\ &= \int_{\mathbb{R}^n} F(\mathbf{k}) e^{2\pi i \mathbf{k} \cdot \mathbf{x}} \left[\int_S H^* \left(s^t \mathbf{k} \right) H \left(s^t \mathbf{k} \right) ds_L \right] d^n k. \end{aligned}$$

Therefore, if the inner integral is essentially bounded as a function of $\mathbf{k}$,

$$\widetilde{F}(\mathbf{k}) = F(\mathbf{k}) \int_S \left| H \left(s^t \mathbf{k} \right) \right|^2 ds_L = F(\mathbf{k}) C_h(\mathbf{k}). \tag{9}$$

Since, by assumption, $C_h(\mathbf{k})$ is in fact a finite constant, the result now follows. ∎

If $C_h(\mathbf{k})$ is essentially bounded, the function h will be said to be S-admissible and the vector space of S-admissible functions will be called $L^2(\mathbb{R}^n, S)$. Note that for a given S it is possible that $C_h(\mathbf{k})$ is infinite for all nonzero $h \in L^2(\mathbb{R}^n)$ and $\mathbf{k} \in \mathbb{R}^n$. In other words, $L^2(\mathbb{R}^n, S)$ may contain only the zero function. Also note that if $C_{h_1}(\mathbf{k})$ and $C_{h_2}(\mathbf{k})$ are constant and $h_3 = h_1 + h_2$, this does not necessarily imply that $C_{h_3}(\mathbf{k})$ is constant. In other words, the set of h such that $C_h(\mathbf{k})$ is a finite constant is not necessarily a vector space. The next proposition shows, however, that it can be, under the right conditions.

PROPOSITION 2. *If the S^t-orbit of a nonzero $\mathbf{k}$ in $\mathbb{R}^n$ is $\mathbb{R}^n \backslash \{0\}$, then $C_h(\mathbf{k})$ is independent of $\mathbf{k}$ for any h and (6) is true when C_h is finite.*

Proof. Suppose $\mathbf{k}_1 \neq \mathbf{k}$ and there is an $s_1 \in S$ such that $s_1^t \mathbf{k} = \mathbf{k}_1$. Then

$$C_h(\mathbf{k}_1) = \int_S \left| H\left(s^t \mathbf{k}_1\right)\right|^2 ds_L = \int_S \left| H\left(s^t s_1^t \mathbf{k}\right)\right|^2 ds_L$$
$$= \int_S \left| H\left(\left(s_1 s\right)^t \mathbf{k}\right)\right|^2 ds_L = C_h(\mathbf{k}).$$

The last equality uses the left invariance of ds_L. In other words, for all $s \in S$,

$$C_h(s^t \mathbf{k}) = C_h(\mathbf{k}).$$

If the S^t-orbit of a nonzero $\mathbf{k}$ in $\mathbb{R}^n$ is $\mathbb{R}^n \backslash \{0\}$, then such an s_1 may be found for any $\mathbf{k}_1 \neq 0$, and therefore $C_h(\mathbf{k})$ is a constant C_h. □

Note that in this case $c_h(\mathbf{x}) = C_h \delta(\mathbf{x})$. It is of course possible for $C_h(\mathbf{k})$ to be a constant for a specific function h even when S^t is not transitive on $\mathbb{R}^n \backslash \{0\}$. In general, however, if h is S-admissible, then $\widetilde{F}(\mathbf{k})$ is the product of $F(\mathbf{k})$ with an S^t-invariant bounded function $C_h(\mathbf{k})$. If $*$ denotes n-dimensional convolution, then this means that

$$\widetilde{f}(\mathbf{x}) = \left(c_h * f\right)(\mathbf{x}), \tag{10}$$

where the distribution c_h satisfies

$$J(s) c_h(s\mathbf{x}) = c_h(\mathbf{x}).$$

In this case the integral in (6) must be followed by a deconvolution to retrieve f from w.

4. Properties of the wavelet transform. The properties of the generalized wavelet transform and its adjoint listed in the next theorem will be useful in discussing the concept of angular wavelet channels. All of them are well known in the more general setting of the theory of square integrable representations of locally compact, nonunimodular groups. The advantage in the presentation below is that all derivations are elementary and the inner product $(h_1, h_2)_S$ is easily computed.

If S^t is transitive on $\mathbb{R}^n \backslash \{0\}$, then we may define an inner product of $h_1, h_2 \in L^2(\mathbb{R}^n, S)$ as

$$(h_1, h_2)_S = \int_S H_1^*\left(s^t \mathbf{k}\right) H_2\left(s^t \mathbf{k}\right) ds_L,$$

which makes this vector space into a Hilbert space. Note that the integral is independent of $\mathbf{k}$. If w is given by (4), write $w = \mathcal{W}\{f, h\}$ and let $\mathcal{W}^\top\{w, h\}$ be the adjoint transform given in (8). If $w_1, w_2 \in L^2(G)$, then let $(w_1, w_2)_G = \int_G w_1^*(g) w_2(g) dg_L$. Let $(f_1, f_2) = \int_{R^n} f_1^*(\mathbf{x}) f_2(\mathbf{x}) d^n x$ be the usual inner product of $f_1, f_2 \in L^2(\mathbb{R}^n)$. Define convolution for group functions in the standard way via $(w_1 * w_2)(g_0) = \int_G w_1(g) w_2(g^{-1} g_0) dg_L$ and let $p_{h_1 h_2} = \mathcal{W}\{h_1, h_2\}$.

PROPOSITION 3. *Suppose that S^t is transitive on $\mathbb{R}^n \backslash \{0\}$; $f, f_1, f_2 \in L^2(\mathbb{R}^n)$ and $h, h_1, h_2 \in L^2(\mathbb{R}^n, S)$. Let $w = \mathcal{W}\{f, h\}, w_1 = \mathcal{W}\{f_1, h_1\}$, and $w_2 = \mathcal{W}\{f_2, h_2\}$. Then $w, w_1, w_2 \in L^2(G)$ and*

$$(w_1, w_2)_G = (f_1, f_2)(h_2, h_1)_S,$$

$$\mathcal{W}^\top\{\mathcal{W}\{f, h_1\}, h_2\} = (h_1, h_2)_S f,$$

$$w * p_{h_1 h_2} = (h, h_1)_S \, \mathcal{W}\{f, h_2\}.$$

Proof. Note that for $g = (s, \mathbf{t})$,

$$w(g) = w_s(\mathbf{t}) = \int_{\mathbb{R}^n} F(\mathbf{k}) H_s^*(\mathbf{k}) e^{-2\pi i \mathbf{k}\cdot\mathbf{t}} d^n k.$$

Then

$$\begin{aligned}
(11) \qquad \int_G w_1^*(g) w_2(g) dg_L &= \int_S \left[\int_{\mathbb{R}^n} w_{1s}^*(\mathbf{t}) w_{2s}(\mathbf{t}) d^n t \right] \frac{1}{J(s)} ds_L \\
&= \int_S \left[\int_{\mathbb{R}^n} F_1^*(\mathbf{k}) H_{1s}(\mathbf{k}) F_2(\mathbf{k}) H_{2s}^*(\mathbf{k}) d^n k \right] \frac{1}{J(s)} ds_L \\
&= \int_{\mathbb{R}^n} F_1^*(\mathbf{k}) F_2(\mathbf{k}) \left[\int_S H_{1s}(\mathbf{k}) H_{2s}^*(\mathbf{k}) \frac{1}{J(s)} ds_L \right] d^n k \\
&= \int_{\mathbb{R}^n} F_1^*(\mathbf{k}) F_2(\mathbf{k}) \left[\int_S H_1\left(s^t \mathbf{k}\right) H_2^*\left(s^t \mathbf{k}\right) ds_L \right] d^n k.
\end{aligned}$$

Setting $w_2 = w_1 = w$ shows that $w \in L^2(G)$. This equation also gives the first relation. For the second, merely repeat the proof of Proposition 1 using h_1 and h_2 instead of h. For the third, let $w'(g) = p_{h_1 h_2}^*(g^{-1} g_0)$ and note that $w' = \mathcal{W}\{\Phi_{g_0} h_2, h_1\}$. If $w'' = W\{f, h_2\}$, then

$$(w * p_{h_1 h_2})(g_0) = (w', w)_G = (\Phi_{g_0} h_2, f)(h, h_1)_S = (h, h_1)_S \, w''(g_0).$$

This gives the result. □

Let $C_{h_1 h_2}(\mathbf{k})$ be the inner integral in (11). In the general case, since h_1 and h_2 are S-admissible, this is a bounded function of $\mathbf{k}$. Therefore the last integral is finite. Then without the orbit condition on S the relations in this proposition are

$$(w_1, w_2)_G = (F_1, C_{h_1 h_2} F_2),$$

$$\mathcal{W}^\top \{\mathcal{W}\{f, h_1\}, h_2\} = c_{h_1 h_2} * f,$$

$$w * p_{h_1 h_2} = \mathcal{W}\{c_{h h_1} * f, h_2\}.$$

5. Orthogonal channels in the wavelet transform. Obviously everything is nicer when S^t is transitive on $\mathbb{R}^n \backslash \{0\}$, which will be the assumption from here on. Let U be the closed subspace of $L^2(G)$ spanned by the set $\{\mathcal{W}\{f, h\} : f \in L^2(\mathbb{R}^n), h \in L^2(\mathbb{R}^n, S)\}$. The left regular representation of G on $L^2(G)$ is given by $L_g w(g_0) = w(g^{-1} g_0)$. It is easy to show that

$$\mathcal{W}\left\{\overline{\Phi}_g f, h\right\} = L_g \mathcal{W}\{f, h\}.$$

Therefore U is a left invariant subspace. Now suppose that $\{h_i\}$ is an orthonormal set in $L^2(\mathbb{R}^n, S)$, i.e., $(h_i, h_j)_S = \delta_{ij}$. If $U_i = \left\{\mathcal{W}\{f, h_i\} : f \in L^2(\mathbb{R}^n)\right\}$, then the U_i are left invariant, mutually orthogonal subspaces of U and $U = \bigoplus U_i$. If $p_{ij} = p_{h_i h_j}$ and $w_i = \mathcal{W}\{f, h_i\}$, then Proposition 3 gives

$$w_i * p_{jk} = \delta_{ij} w_k,$$

$$\mathcal{W}^{\top}\{w_i, h_j\} = \delta_{ij} f.$$

Now let $h = \sum_i a_i h_i$ be an admissible function, and $w = \mathcal{W}\{f, h\}$. Then $w = \sum_i a_i^* w_i$ expresses w as an orthogonal sum in $L^2(G)$. Each w_i is f filtered through the channel h_i. Any w_i may be recovered from w via $w * p_{ii} = a_i^* w_i$. The original function f may be recovered from any w_i via $f = \mathcal{W}^{\top}\{w_i, h_i\} = \frac{1}{a_i}\mathcal{W}^{\top}\{w_i, h\}$. Therefore each w_i contains all of the information about f, but this information may be contained in subtle variations in $w_i(g)$ as g varies in G. If there is noise present these subtle variations may be indiscernible. But, because of the orthogonality conditions, it is possible to compute all of the w_i simultaneously by computing w and then examining them separately by convolving w with functions p_{ii}. If each w_i produces strong responses to different features of f, this could be an efficient way to analyze the function in the presence of noise. Note that $C_h = \sum_i |a_i|^2$.

Operators on $L^2(\mathbb{R}^n)$ may also be analyzed for their effect on the channels. Let $\mathcal{Q}$ be an integral operator on $L^2(\mathbb{R}^n)$ given by $\mathcal{Q}f(\mathbf{x}) = \int_{\mathbb{R}^n} q(\mathbf{x}, \mathbf{y}) f(\mathbf{y}) d^n y$. If $w = \mathcal{W}\{f, h\}$ and $\mathcal{Q}_h w = \mathcal{W}\{\mathcal{Q}f, h\}$, then

$$\mathcal{Q}_h w(g_0) = \int_G q_h(g_0, g) w(g) dg_L,$$

where

$$q_h(g_0, g) = \frac{1}{C_h} \int_{\mathbb{R}^n} \left[\int_{\mathbb{R}^n} q(\mathbf{x}, \mathbf{y}) \overline{\Phi}_{g_0} h^*(\mathbf{x}) d^n x \right] \overline{\Phi}_g h(\mathbf{y}) d^n y. \tag{12}$$

Now let

$$q_{ij}(g_0, g) = \int_{\mathbb{R}^n} \left[\int_{\mathbb{R}^n} q(\mathbf{x}, \mathbf{y}) \overline{\Phi}_{g_0} h_i^*(\mathbf{x}) d^n x \right] \overline{\Phi}_g h_j(\mathbf{y}) d^n y$$

and notice that, for fixed g_0, this is a function in U_j^*. Also

$$q_h(g_0, g) = \frac{1}{C_h} \sum_i \sum_j a_i^* a_j q_{ij}(g_0, g).$$

As above, $w = \sum_k a_k^* w_k$ with $w_k = \mathcal{W}\{f, h_k\}$ and, using the fact that U_j and U_k are orthogonal subspaces when $k \neq j$,

$$\mathcal{Q}_h w(g_0) = \sum_i a_i^* \left[\frac{1}{C_h} \sum_j |a_j|^2 \int_G q_{ij}(g_0, g) w_j(g) dg_L \right].$$

For a fixed g, $q_{ij}(g_0, g)$ is a function in U_i. Since this is a closed subspace, the expression in brackets is also a function in U_i. Therefore $\mathcal{Q}_h w = \sum_i a_i^* \mathcal{W}\{\mathcal{Q}f, h_i\} = \sum_i a_i^* (\mathcal{Q}_h w)_i$ and

$$(\mathcal{Q}_h w)_i(g_0) = \frac{1}{C_h} \sum_j |a_j|^2 \int_G q_{ij}(g_0, g) w_j(g) dg_L.$$

The function q_{ij} then measures how much the operator mixes channel h_j with channel h_i.

6. Angular channels in a two-dimensional wavelet transform. As an example of a channel decomposition consider the $n = 2$ case. An element f of $L^2\left(\mathbb{R}^2\right)$ may be thought of as an image and the wavelet transform could be used for image analysis or modification. Let the point group S be $\mathbb{R}^+ \times SO(2)$, where $\mathbb{R}^+$ is the multiplicative group of positive real numbers. If $s \in S$, then $s = aR(\phi)$, indicating a scale transformation by a factor of a and a rotation by an angle ϕ about the origin. Here $R(\phi)$ is the usual 2×2 rotation matrix associated with ϕ,

$$R(\phi) = \begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix}.$$

If $s' = a'R(\phi')$, then $ss' = aa'R(\phi + \phi')$. The invariant measures are

$$ds_L = \left(\frac{da}{a}\right) d\phi,$$

$$dg_L = \left(\frac{d^2t}{J(s)}\right) ds_L = \frac{da}{a^3} d\phi d^2t.$$

Let J be the interval $[0, 2\pi)$. The detailed wavelet formulae for this case are

$$w(a, \phi, \mathbf{t}) = \int_{\mathbb{R}^2} f(\mathbf{x}) \frac{1}{a} h^* \left(\frac{1}{a} R(-\phi)(\mathbf{x} - \mathbf{t})\right) d^2x,$$

$$f(\mathbf{x}) = \frac{1}{C_h} \int_{\mathbb{R}^2} \int_J \int_{\mathbb{R}^+} w(a, \phi, \mathbf{t}) \frac{1}{a} h \left(\frac{1}{a} R(-\phi)(\mathbf{x} - \mathbf{t})\right) \frac{da}{a^3} d\phi d^2t,$$

$$C_h = \int_J \int_{\mathbb{R}^+} \left|H\left(aR(-\phi)\mathbf{k}\right)\right|^2 \frac{da d\phi}{a} < \infty.$$

This admissibility condition is equivalent to

$$C_h = \int_{\mathbb{R}^2} \left|H\left(\mathbf{k}\right)\right|^2 \frac{d^2k}{\left|\mathbf{k}\right|^2} < \infty.$$

Similarly, the wavelet inner product is given by

$$\left(h_1, h_2\right)_S = \int_{\mathbb{R}^2} H_1^*\left(\mathbf{k}\right) H_2\left(\mathbf{k}\right) \frac{d^2k}{\left|\mathbf{k}\right|^2}.$$

The channel expansion for h will be the Fourier expansion in the angular coordinate. This is also the expansion of h into basis functions for irreducible representations of $SO(2)$. Let $\theta(\widehat{\mathbf{x}})$ be the polar angle of the vector $\mathbf{x}$. Then

$$h(\mathbf{x}) = \sum_{n=-\infty}^{\infty} a_n \overline{h}_n\left(|\mathbf{x}|\right) \exp[in\theta\left(\widehat{\mathbf{x}}\right)] = \sum_{n=-\infty}^{\infty} a_n h_n\left(\mathbf{x}\right),$$

where

$$\overline{h}_n\left(r\right) = \frac{1}{2\pi a_n} \int_J h(r, \theta) e^{-in\theta} d\theta,$$

and a_n is chosen so that $(h_n, h_n)_S = 1$. If we let

$$\overline{H}_n(k) = 2\pi \int_{\mathbb{R}^+} J_n(2\pi kr)\overline{h}_n(r) r dr,$$

where J_n is the nth order Bessel function of the first kind, then

$$H_n(\mathbf{k}) = \overline{H}_n\left(|\mathbf{k}|\right) \exp\left[in\theta\left(\widehat{\mathbf{k}}\right)\right].$$

This implies that $(h_n, h_m)_S = \delta_{nm.}$ The normalization condition on h_n is equivalent to

$$2\pi \int_{\mathbb{R}^+} \left|\overline{H}_n\left(a\,|\mathbf{k}|\right)\right|^2 \frac{da}{a} = 2\pi \int_{\mathbb{R}^+} \left|\overline{H}_n(a)\right|^2 \frac{da}{a} = 1.$$

Scaling and rotating h_n gives

$$\begin{aligned} h_n\left(\frac{1}{a}R(-\phi)\mathbf{x}\right) &= \overline{h}_n\left(\frac{1}{a}\,|\mathbf{x}|\right) \exp\left[in(\theta\left(\widehat{\mathbf{x}}\right) - \phi)\right] \\ &= \exp(-in\phi) h_n\left(\frac{1}{a}\mathbf{x}\right). \end{aligned}$$

The forward transform now has the form

$$w(a, \phi, \mathbf{t}) = \sum_{n=-\infty}^{\infty} a_n^* \exp(in\phi)\overline{w}_n\left(a, \mathbf{t}\right), \tag{13}$$

with

$$\overline{w}_n\left(a, \mathbf{t}\right) = \int_{\mathbb{R}^2} f(\mathbf{x}) \frac{1}{a} h_n^*\left(\frac{1}{a}(\mathbf{x} - \mathbf{t})\right) d^2x.$$

Therefore the order n wavelet channel is given by

$$w_n\left(a, \phi, \mathbf{t}\right) = \exp(in\phi)\overline{w}_n\left(a, \mathbf{t}\right).$$

Using w_n and h_n to recover f results in

$$f(\mathbf{x}) = 2\pi \int_{\mathbb{R}^2} \int_{\mathbb{R}^+} \overline{w}_n\left(a, \mathbf{t}\right) \frac{1}{a} h_n\left(\frac{1}{a}(\mathbf{x} - \mathbf{t})\right) \frac{da}{a^3} d^2t. \tag{14}$$

Notice that each $\overline{w}_n\left(a, \mathbf{t}\right)$ is a wavelet transform with point group $\{aI : a > 0\}$ and wavelet h_n and that (14) is the corresponding inversion formula. This point group, in contrast to $\mathbb{R}^+ \times SO(2)$, is not transitive on $\mathbb{R}^2 \backslash \{0\}$ but $C_{h_n}(\mathbf{k})$ is still a constant due to the particular angular dependence of h_n. Thus each angular channel h_n gives rise via scaling and translations to a wavelet transform $w_n(a, \phi, \mathbf{t})$ and each transform contains all of the information about f. But each channel will have strong responses to different local angular features of the image. To examine these features the projection functions may be convolved with w. They are given by

$$p_{nm}\left(a, \phi, \mathbf{t}\right) = e^{im\phi}\overline{p}_{nm}\left(a, \mathbf{t}\right),$$

with

$$\overline{p}_{nm}\left(a, \mathbf{t}\right) = \int_{\mathbb{R}^2} h_n\left(\mathbf{x}\right) \frac{1}{a} h_m^*\left(\frac{1}{a}\left(\mathbf{x} - \mathbf{t}\right)\right) d^2x.$$

From $w_n = w_m * p_{mn}$, we have

$$\overline{w}_n(a_0, \mathbf{t}_0) = \int_{\mathbb{R}^2} \int_{\mathbb{R}^+} \overline{w}_m(a, \mathbf{t}) \widehat{p}_{mn}(a, \mathbf{t}; a_0, \mathbf{t}_0) \frac{1}{a^3} da d^2t,$$

with

$$\widehat{p}_{mn}(a, \mathbf{t}; a_0, \mathbf{t}_0) = \int_J \overline{p}_{mn} \left(\frac{a_0}{a}, \frac{1}{a} R(-\phi)\mathbf{t}_0 - \mathbf{t} \right) e^{i(m-n)\phi} d\phi.$$

This looks rather complicated. Fortunately there is an easier way:

$$\overline{w}_n(a, \mathbf{t}) = \frac{1}{2\pi} \int_J w(a, \phi, \mathbf{t}) e^{-in\phi} d\phi,$$

which is actually more useful since this operation could be performed numerically with the fast Fourier transform. The prescription then is to compute w for a given f, and to Fourier transform the angular variable in w to pick out the response w_n of each channel.

For operators, we have

$$q_{nm}(a_0, \phi_0, \mathbf{t}_0; a, \phi, \mathbf{t}) = e^{i(n\phi_0 - m\phi)} \overline{q}_{nm}(a_0, \mathbf{t}_0; a, \mathbf{t}),$$

with

$$\overline{q}_{nm}(a_0, \mathbf{t}_0; a, \mathbf{t}) = \int_{\mathbb{R}^2} \int_{\mathbb{R}^2} q(\mathbf{x}, \mathbf{y}) \frac{1}{a_0} h_n^* \left(\frac{1}{a_0} (\mathbf{x} - \mathbf{t}_0) \right) \frac{1}{a} h_m \left(\frac{1}{a} (\mathbf{x} - \mathbf{t}) \right) d^2x d^2y$$

and

$$(\mathcal{Q}_h w)_n(a_0, \phi_0, \mathbf{t}_0) = e^{in\phi_0} \sum_{m=-\infty}^{\infty} |a_m|^2 \int_{\mathbb{R}^2} \int_{\mathbb{R}^+} \overline{q}_{nm}(a_0, \mathbf{t}_0; a, \mathbf{t}) \overline{w}_m(a, \mathbf{t}) \frac{da}{a^3} d^2t.$$

7. An example. Let $\overline{h}_0(r) = 0$, and for $n \neq 0$ let $\overline{h}_n(r) = b_n$ when $r < 1$ and zero otherwise. Also let $a_n = \frac{\lambda^n}{b_n}$ with $0 < \lambda < 1$. In other words, $h(r, \theta) = \mathrm{Re}[\sum_{n=1}^{\infty} \lambda^n e^{in\theta}]$ for $r < 1$. The following function $\overline{H}_n(k)$ may be calculated in this case:

$$\overline{H}_n(k) = \frac{2b_n}{\pi k^2} \left[\pi k \sum_{j=0}^{\infty} J_{n+2j+1}(2\pi k) - \sum_{j=0}^{\infty} (j+1) J_{n+2j+2}(2\pi k) \right].$$

For small k, $J_n(2\pi k) \sim \frac{(\pi k)^n}{(n+1)!}$ and

$$\overline{H}_n(k) \sim 2\pi b_n (\pi k)^n.$$

Thus each h_n is admissible. Explicitly, for $r < 1$,

$$h(r, \theta) = \frac{2\lambda [\cos(\theta) - \lambda]}{(1 + \lambda^2) - 2\lambda \cos(\theta)}.$$

For a real image f, the function $\mathrm{Re}[w_1]$ will have a strong peak when $(a_0, \phi_0, \mathbf{t}_0)$ localizes h_1 with its center on an edge oriented in direction $\phi_0 + \frac{\pi}{2}$ and longer than $2a_0$. The function $\mathrm{Re}[w_2]$ will have a strong peak when $(a_0, \phi_0, \mathbf{t}_0)$ places the center

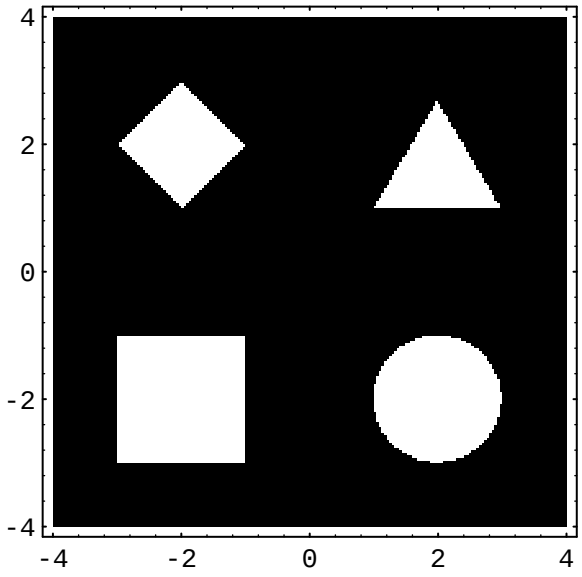


FIG. 1. *The function f.*

of h_2 on a right angled corner oriented with an edge at angle $\phi_0 + \frac{\pi}{4}$. Other orders in h have a similar behavior. In a noisy system these strong peaks are desirable for analyzing an image.

To illustrate some of these points Figure 1 shows a binary function f. In Figures 2–6 the functions $\mathrm{Re}[w_1(a,\phi,\mathbf{t})]$ are shown as functions of $\mathbf{t} = (t_1, t_2)$ with $a = .25$ and $\phi = 0,\ \frac{\pi}{6},\ \frac{\pi}{4},\ \frac{\pi}{3},\ \frac{\pi}{2}$, respectively. As expected this channel responds strongly to edges oriented perpendicular to ϕ and gives zero response when an edge is oriented parallel to ϕ. In Figures 7–10 are similar plots of $\mathrm{Re}[w_2(a,\phi,\mathbf{t})]$ for $\phi = 0,\ \frac{\pi}{6},\ \frac{\pi}{4},\ \frac{\pi}{3}$. This channel has extreme points at locations where there is a right angle oriented with its bisector perpendicular or parallel to ϕ. On edges parallel or perpendicular to ϕ it produces a peak and a trough next to each other. For Figures 11–13 $\mathrm{Re}[w_3(a,\phi,\mathbf{t})]$ is shown for $\phi = 0,\ \frac{\pi}{6},\ \frac{\pi}{4}$. Finally in Figures 14–18 $\mathrm{Re}[w(a,\phi,t)]$ is plotted for $\phi = 0,\ \frac{\pi}{6},\ \frac{\pi}{4},\ \frac{\pi}{3},\ \frac{\pi}{2}$ and $\lambda = \frac{1}{2}$. This function is the sum of $\mathrm{Re}[\lambda^n w_n]$ for $n = 1, 2, 3, \ldots$. Notice that the corner detecting ability of w_2 is hidden when the sum is taken. Of course any w_n can be retrieved from w by Fourier transforming in the ϕ variable and evaluating at frequency n.

8. Angular channels in a three-dimensional wavelet transform. Now let $n = 3$ so that f may be considered to be some three-dimensional distribution of intensity or other interesting quantity. Take $S = \mathbb{R}^+ \times SO(3)$ and $T = \mathbb{R}^3$. For $\sigma \in SO(3)$, the wavelet formulae now look like

$$w(a,\sigma,\mathbf{t}) = \int_{\mathbb{R}^3} f(\mathbf{x}) \frac{1}{\sqrt{a^3}} h^* \left(\frac{1}{a} \sigma^{-1}(\mathbf{x} - \mathbf{t}) \right) d^2x,$$

$$f(\mathbf{x}) = \frac{1}{C_h} \int_{\mathbb{R}^3} \int_{SO(3)} \int_{\mathbb{R}^+} w(a,\sigma,\mathbf{t}) \frac{1}{\sqrt{a^3}} h \left(\frac{1}{a} \sigma^{-1}(\mathbf{x} - \mathbf{t}) \right) \frac{da}{a^4} d\sigma_L d^3t,$$

$$C_h = \int_{SO(3)} \int_{\mathbb{R}^+} \left| H\left(a\sigma^t \mathbf{k}\right) \right|^2 \frac{da}{a} d\sigma_L < \infty.$$

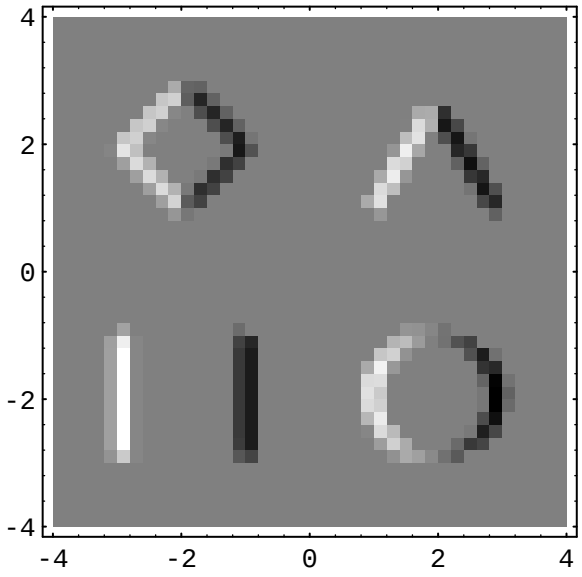


FIG. 2. *Real part of $n = 1$ component of w at angle 0 degrees.*

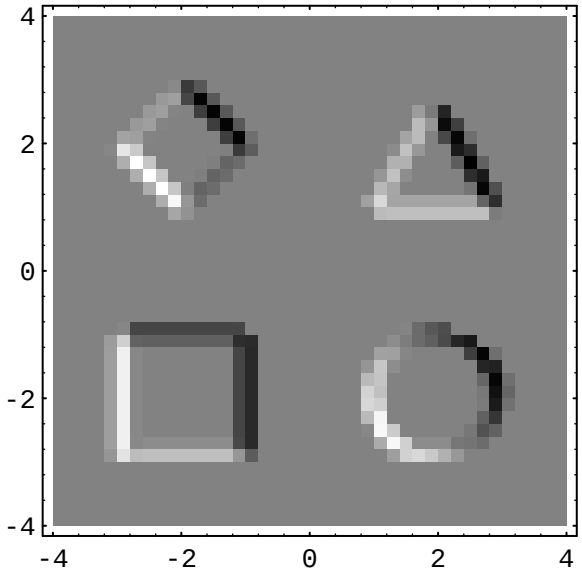


FIG. 3. *Real part of $n = 1$ component of w at angle 30 degrees.*

If $d\sigma_L$ is suitably normalized, it can be shown that the admissibility condition is

$$C_h = \int_{\mathbb{R}^3} |H(\mathbf{k})|^2 \frac{d^3k}{|\mathbf{k}|^3} < \infty.$$

Similarly, the wavelet inner product is

$$(h_1, h_2)_S = \int_{\mathbb{R}^3} H_1^*(\mathbf{k}) H_2(\mathbf{k}) \frac{d^3k}{|\mathbf{k}|^3}.$$

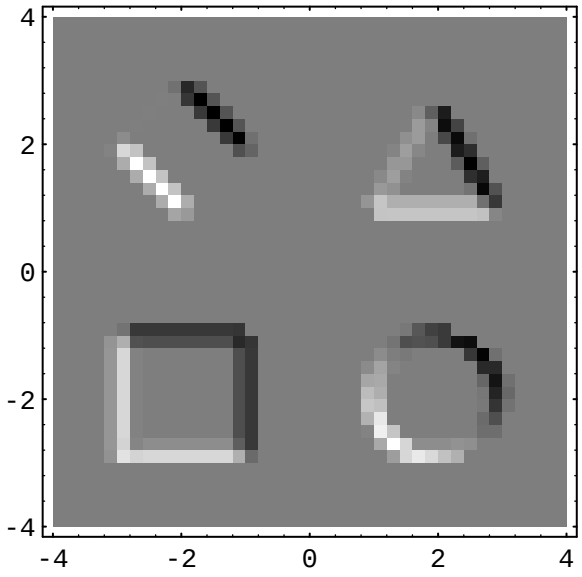


FIG. 4. *Real part of $n = 1$ component of w at angle 45 degrees.*

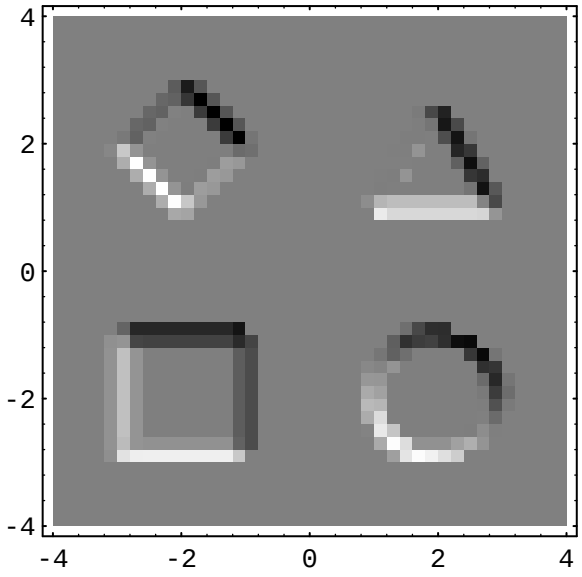


FIG. 5. *Real part of $n = 1$ component of w at angle 60 degrees.*

The channel situation is more complicated. The angular decomposition of h is given by

$$h(\mathbf{x}) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} a_m^l \overline{h}_m^l \left(|\mathbf{x}|\right) \Psi_m^l \left(\widehat{\mathbf{x}}\right) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} a_m^l h_m^l(\mathbf{x}),$$

where the functions Ψ_m^l are the spherical harmonics. For a given l, they form a basis for an irreducible unitary representation of $SO(3)$ of dimension $2l + 1$. This means

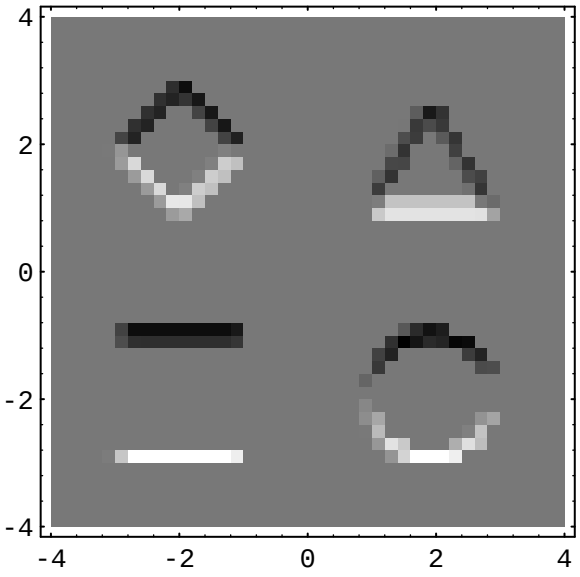


FIG. 6. *Real part of $n = 1$ component of w at angle 90 degrees.*

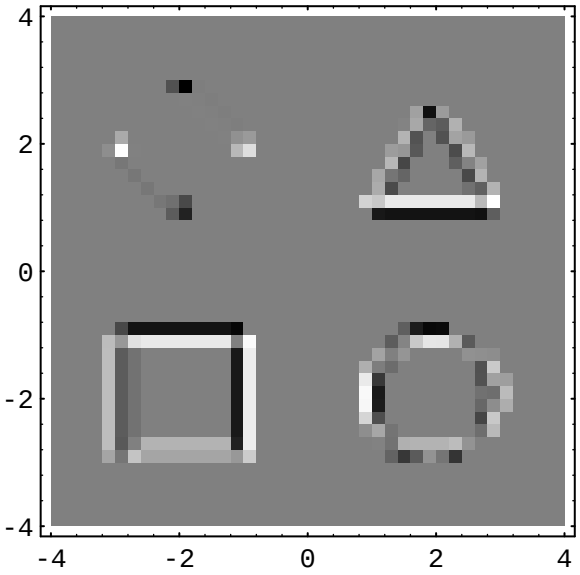


FIG. 7. *Real part of $n = 2$ component of w at angle 0 degrees.*

that

$$\Psi_m^l\left(\sigma^{-1}\widehat{\mathbf{x}}\right) = \sum_{n=-l}^{l} U_{nm}^l(\sigma)\,\Psi_n^l(\widehat{\mathbf{x}}),$$

with the matrices $U^l(\sigma)$ determined by the corresponding representation. In Fourier space

$$H_m^l(\mathbf{k}) = \overline{H}_m^l(|\mathbf{k}|)\,\Psi_m^l\left(\widehat{\mathbf{k}}\right),$$

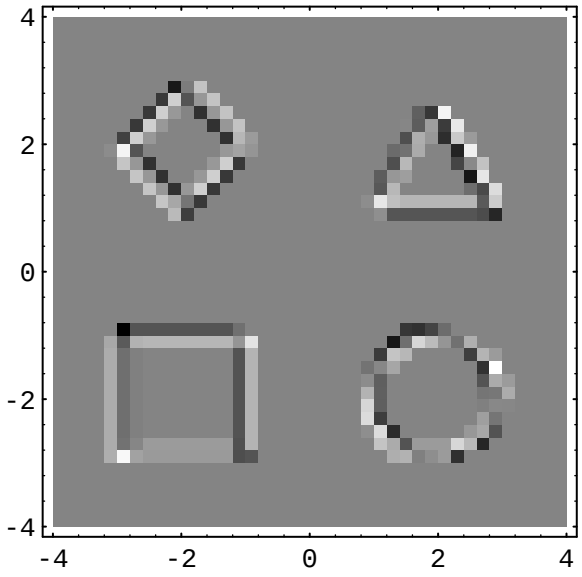


FIG. 8. *Real part of $n = 2$ component of w at angle 30 degrees.*

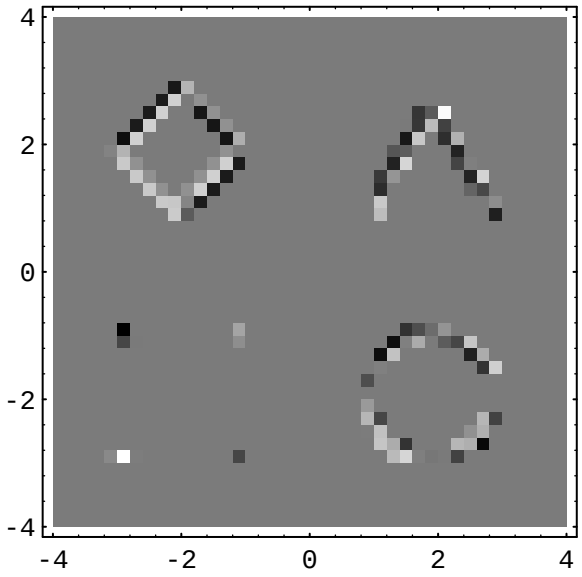


FIG. 9. *Real part of $n = 2$ component of w at angle 45 degrees.*

where

$$\overline{H}_m^l(k) = 4\pi i^l \int_{\mathbb{R}^+} \overline{h}_m^l(r) j_l(2\pi kr) r^2 dr$$

and j_l is the degree l spherical Bessel function of the first kind. Therefore if $l \neq l'$ or $m \neq m'$, then $\left(h_m^l, h_{m'}^{l'}\right)_S = 0$. As before, choose the a_m^l so that $\left(h_m^l, h_m^l\right)_S = 1$.

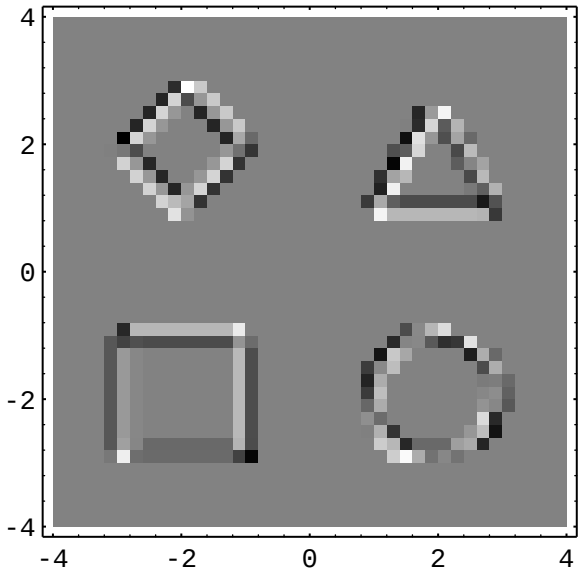


FIG. 10. *Real part of $n = 2$ component of w at angle 60 degrees.*

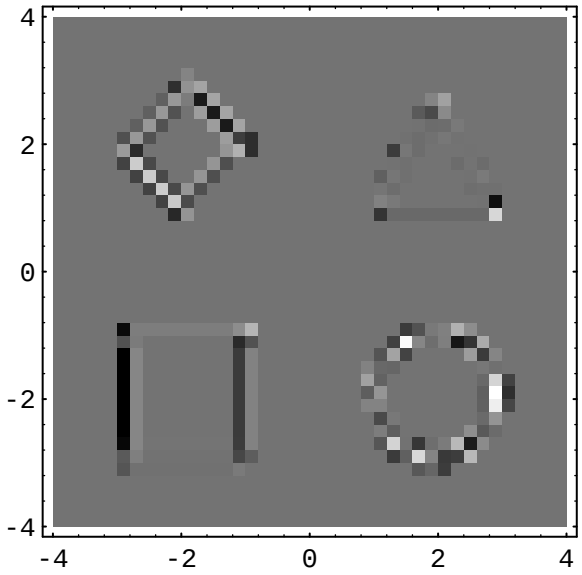


FIG. 11. *Real part of $n = 3$ component of w at angle 0 degrees.*

For the wavelet transform, let $h^l_{mn}(\mathbf{x}) = \overline{h}^l_m(|\mathbf{x}|)\,\Psi^l_n(\widehat{\mathbf{x}})$. Then rotating and scaling h^l_m gives

$$h^l_m\left(\frac{1}{a}\sigma^{-1}\mathbf{x}\right) = \overline{h}^l_m\left(\frac{1}{a}|\mathbf{x}|\right)\Psi^l_m\left(\sigma^{-1}\widehat{\mathbf{x}}\right) = \sum_{n=-l}^{l} U^l_{nm}(\sigma)\,h^l_{mn}(\mathbf{x}).$$

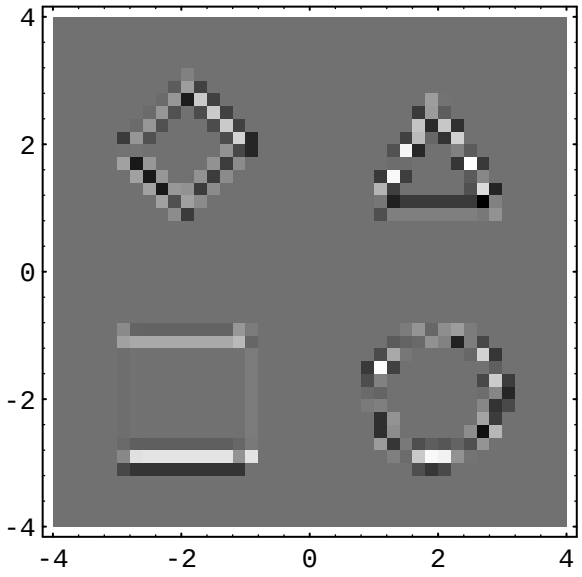


FIG. 12. *Real part of $n = 3$ component of w at angle 30 degrees.*

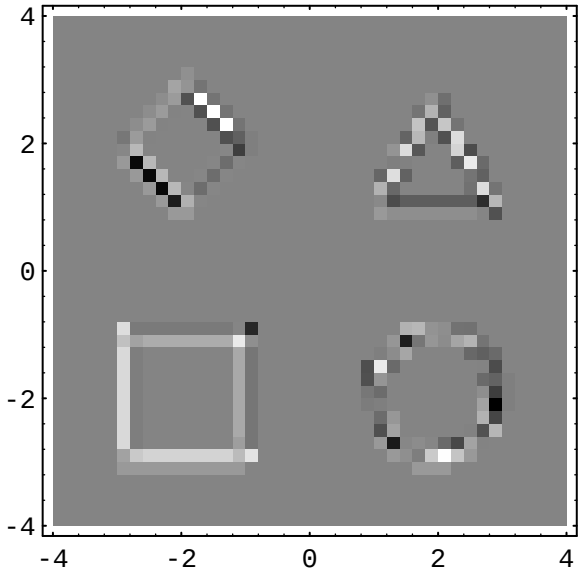


FIG. 13. *Real part of $n = 3$ component of w at angle 60 degrees.*

Putting this into the forward transform gives the response in the angular channel h_m^l:

$$w_m^l(a, \sigma, \mathbf{t}) = \sum_{n=-l}^{l} \left[U_{nm}^l(\sigma)\right]^* \overline{w}_{mn}^l(a, \mathbf{t}),$$

with

$$\overline{w}_{mn}^l(a, \mathbf{t}) = \int_{\mathbb{R}^3} f(\mathbf{x}) \frac{1}{\sqrt{a^3}} \left[h_{mn}^l\left(\frac{1}{a}(\mathbf{x} - \mathbf{t})\right)\right]^* d^3x.$$

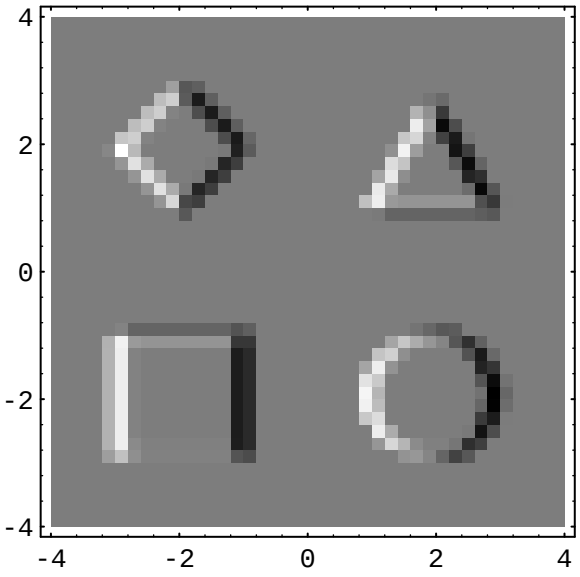


FIG. 14. *Real part of w at angle* 0 *degrees.*

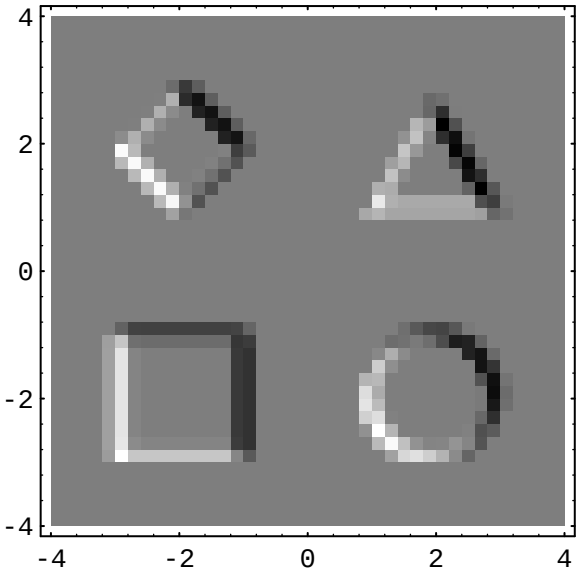


FIG. 15. *Real part of w at angle* 30 *degrees.*

Notice again that this is a scale only wavelet transform of f. This suggests the definition

$$w^l_{mn}(a,\sigma,\mathbf{t}) = \left[U^l_{nm}(\sigma)\right]^* \overline{w}^l_{mn}(a,\mathbf{t}),$$

which could be called the nth angular subchannel of w^l_m.

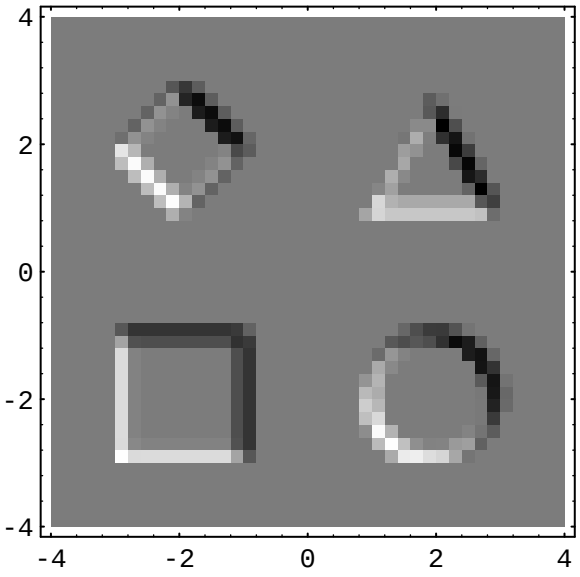


FIG. 16. *Real part of w at angle 45 degrees.*

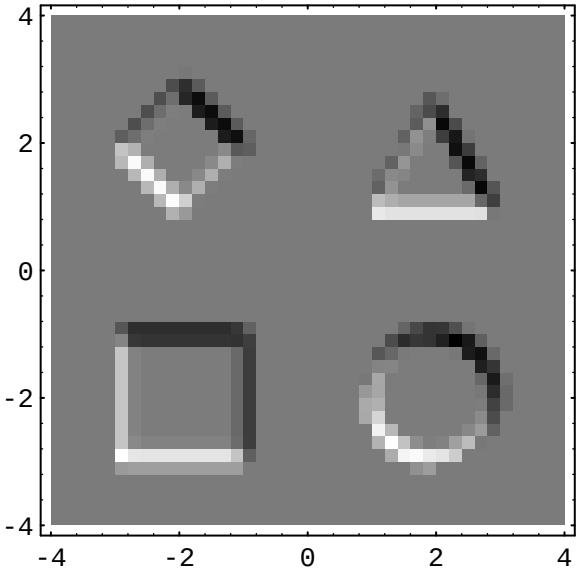


FIG. 17. *Real part of w at angle 60 degrees.*

When f is recovered from w_m^l and h_m^l, the result is

$$\begin{aligned} f(\mathbf{x}) &= \int_{\mathbb{R}^3}\int_{SO(3)}\int_{\mathbb{R}^+} w_m^l(a,\sigma,\mathbf{t})\frac{1}{\sqrt{a^3}}h_m^l\left(\frac{1}{a}\sigma^{-1}\left(\mathbf{x}-\mathbf{t}\right)\right)\frac{da}{a^4}d\sigma_L d^3t \\ &= \int_{\mathbb{R}^3}\int_{SO(3)}\int_{\mathbb{R}^+}\sum_{n=-l}^{l}\left[U_{nm}^l\left(\sigma\right)\right]^*\overline{w}_{mn}^l(a,\mathbf{t})\frac{1}{\sqrt{a^3}} \end{aligned}$$

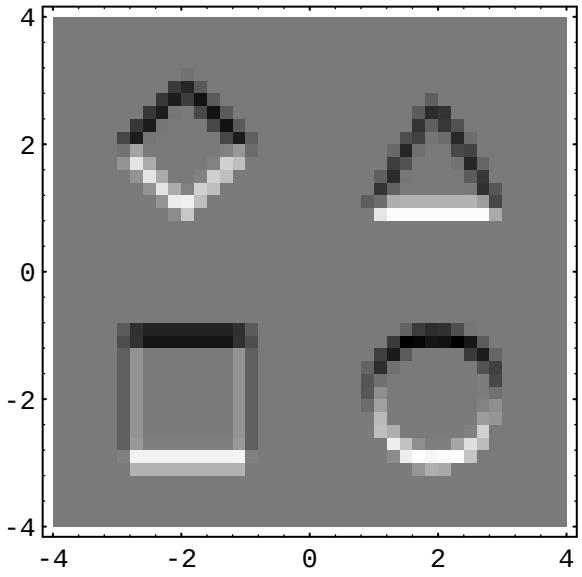


FIG. 18. *Real part of w at angle 90 degrees.*

$$\cdot \left[\sum_{n'=-l}^{l} U^l_{n'm}(\sigma)\, h^l_{mn'} \left(\frac{1}{a}(\mathbf{x}-\mathbf{t}) \right) \right] \frac{da}{a^4} d\sigma_L d^3t.$$

However, a fact from representation theory is that

$$\int_{SO(3)} \left[U^l_{nm}(\sigma) \right]^* U^{l'}_{n'm'}(\sigma) d\sigma_L = \frac{\omega}{2l+1} \delta_{ll'} \delta_{mm'} \delta_{nn'}, \tag{15}$$

where $\int_{SO(3)} d\sigma_L = \omega$. With the present normalization for $d\sigma_L$, $\omega = 4\pi$, so

$$f(\mathbf{x}) = \frac{4\pi}{2l+1} \sum_{n=-l}^{l} \int_{\mathbb{R}^3} \int_{\mathbb{R}^+} \overline{w}^l_{mn}(a,\mathbf{t}) \frac{1}{\sqrt{a^3}} h^l_{mn} \left(\frac{1}{a}(\mathbf{x}-\mathbf{t}) \right) \frac{da}{a^4} d^3t.$$

Notice that the subchannels h^l_{mn} do not interfere when reconstructing f from w^l_m. Normally, in a data compression or filtering operation, w would be sampled on some discrete set in G, which is eight dimensional, and these samples would be modified in some way. Then f would be reconstructed by discretizing the inverse transform on the same subset of G. However, if the harmonic expansion of h is known, then the $\overline{w}^l_{mn}$ could be sampled on $\mathbb{R}^+ \times \mathbb{R}^3$, a four-dimensional space, for a finite set of (l,m,n). After the samples are modified, this last integral could be discretized on the same set for reconstruction. If h is well approximated by a small number of its angular components, this could be a more efficient process.

Equation (15) also implies that

$$\left(w^l_{mn}, w^{l'}_{m'n'} \right)_G = 0$$

when $l \neq l'$, $m \neq m'$, or $n \neq n'$. Thus the subchannels may be projected from w:

$$\overline{w}^l_{mn}(a,\mathbf{t}) = \frac{(2l+1)}{\omega} \int_{SO(3)} U^l_{nm}(\sigma)\, w(a,\sigma,\mathbf{t}) d\sigma_L.$$

However, f cannot in general be reconstructed from one subchannel transform without a deconvolution step. Again, each channel and subchannel provides different local angular information about f.

For projection functions, we have

$$p_{mm'}^{ll'}(a,\sigma,\mathbf{t}) = \sum_{n'=-l}^{l} U_{n'm'}^{l'}(\sigma)^* \, \overline{p}_{mm',nn'}^{ll'}(a,\mathbf{t}),$$

with

$$\overline{p}_{mm',nn'}^{ll'}(a,\mathbf{t}) = \int_{\mathbb{R}^3} h_{mn}^{l}(\mathbf{x}) \frac{1}{\sqrt{a^3}} \left[h_{m'n'}^{l'}\left(\frac{1}{a}(\mathbf{x}-\mathbf{t})\right)\right]^* d^3x.$$

For an operator,

$$q_{mm'}^{ll'}(a_0,\sigma_0,\mathbf{t}_0;a,\sigma,\mathbf{t}) = \sum_{n=-l}^{l}\sum_{n'=-l}^{l} U_{nm}^{l}(\sigma_0)^* \, U_{n'm'}^{l'}(\sigma) \, \overline{q}_{mm',nn'}^{ll'}(a_0\mathbf{t}_0;a,\mathbf{t}),$$

with

$$\begin{aligned}\overline{q}_{mm',nn'}^{ll'}(a_0\mathbf{t}_0;a,\mathbf{t}) = \int_{\mathbb{R}^3}\int_{\mathbb{R}^3} q(\mathbf{x},\mathbf{y}) \frac{1}{\sqrt{a_0^3}} \left[h_{mn}^{l}\left(\frac{1}{a_0}(\mathbf{x}-\mathbf{t}_0)\right)\right]^* \\ \cdot \frac{1}{\sqrt{a^3}} h_{m'n'}^{l'}\left(\frac{1}{a}(\mathbf{y}-\mathbf{t})\right) d^3x d^3y.\end{aligned}$$

Notice that, as in the two-dimensional case, all angular integrations, i.e., those over $SO(3)$, can be removed if the angular decomposition of h is known.

9. Angular channels in higher dimensions and in the Fourier domain. The generalization to a wavelet transform for $L^2(\mathbb{R}^n)$ using $S = \mathbb{R}^+ \times SO(n)$ is straightforward. As before, if $d\sigma_L$ is suitably normalized,

$$(h_1,h_2)_S = \int_{\mathbb{R}^n} H_1^*(\mathbf{k}) \, H_2(\mathbf{k}) \, \frac{d^n k}{|\mathbf{k}|^n}.$$

No new complications arise when decomposing a particular wavelet into angular channels using the appropriate spherical harmonics. Again, integration over $SO(n)$ can be eliminated in the reconstruction of f by using angular subchannels as in the three-dimensional case. Since the dimension of $SO(n)$ is $\frac{n(n-1)}{2}$, sampling advantages increase with dimension if the angular components of h are known.

Another way to avoid integration over $SO(n)$ is to go into Fourier space. In general,

$$W(s,\mathbf{k}) = W_s(\mathbf{k}) = F(\mathbf{k}) H^*(s^t\mathbf{k}) \sqrt{J(s)}.$$

Let $\mathbf{k}_0 = s^t\mathbf{k}$. If $s_1^t\mathbf{k} = s_2^t\mathbf{k}$, then $(s_1^t)^{-1} s_2^t\mathbf{k} = \mathbf{k}$. This implies that $(s_1^t)^{-1} s_2^t \in SO(n)$ and therefore $J(s_1) = J(s_2)$. Then there is a well-defined function $\widehat{W}(\mathbf{k}_0,\mathbf{k})$ satisfying

$$W(s,\mathbf{k}) = \widehat{W}\left(s^t\mathbf{k},\mathbf{k}\right)$$

and since $J(s) = \left(\frac{|\mathbf{k}_0|}{|\mathbf{k}|}\right)^n$,

$$\widehat{W}(\mathbf{k}_0, \mathbf{k}) = F(\mathbf{k}) \left(\frac{|\mathbf{k}_0|}{|\mathbf{k}|}\right)^{\frac{n}{2}} H(\mathbf{k}_0).$$

The forward transform may now be described as follows:

1. Fourier transform $f(\mathbf{x})$ to get $F(\mathbf{k})$.
2. Multiply $F(\mathbf{k})$ by $(\frac{|\mathbf{k}_0|}{|\mathbf{k}|})^{\frac{n}{2}} H(\mathbf{k}_0)$ to get $\widehat{W}(\mathbf{k}_0, \mathbf{k})$.
3. Evaluate $\widehat{W}(\mathbf{k}_0, \mathbf{k})$ on the twisted diagonal $\mathbf{k}_0 = s^t \mathbf{k}$ to get $W(s, \mathbf{k})$.
4. Inverse Fourier transform $W(s, \mathbf{k})$ to get $w(s, \mathbf{t})$.

Let $s^t_{\mathbf{k}_0\mathbf{k}}$ be the element of S corresponding to a rotation and dilation in the plane spanned by $\mathbf{k}$ and $\mathbf{k}_0$ which takes $\mathbf{k}$ to $\mathbf{k}_0$. For a normalized wavelet, the inverse transform can be arrived at via the following:

1. Fourier transform $w(s, \mathbf{t})$ to get $W(s, \mathbf{k})$.
2. Let $\widehat{W}(\mathbf{k}_0, \mathbf{k}) = W(s_{\mathbf{k}_0\mathbf{k}}, \mathbf{k})$.
3. Multiply $\widehat{W}(\mathbf{k}_0, \mathbf{k})$ by $|\mathbf{k}|^{\frac{n}{2}} |\mathbf{k}_0|^{-\frac{3n}{2}} H(\mathbf{k}_0)$ and integrate over $\mathbf{k}_0$ to get $F(\mathbf{k})$.
4. Inverse Fourier transform $F(\mathbf{k})$ to get $f(\mathbf{x})$.

The integration over $SO(n)$ has been replaced by the integration over $\mathbb{R}^n$ in step 3. The spherical harmonic expansion of h gives rise to one for H as before. This corresponds now to the spherical harmonic expansion for $\widehat{W}$ in the first variable. Therefore each w^l_m corresponds to a $\widehat{W}^l_m$. If we expanded f also, this would give the expansion for $\widehat{W}$ in the second variable.

10. Conclusion. Multidimensional wavelet transforms and inverse transforms have been developed which can be used with a wide variety of point symmetry groups. The admissibility condition of a particular wavelet is the boundedness of a certain integral over the point group. The decomposition of a wavelet transform into orthogonal channels has been described and the expression of integral operators in terms of these channels has been given. As an example, angular wavelet channels in two and three dimensions have been developed in detail. In three dimensions, results from the theory of group representations came into play in the notion of angular subchannels. The generalization of angular channels to higher dimensions has been indicated and the implications in the Fourier domain have been outlined.

Acknowledgments. I would like to thank Harry Barrett and Jack Denny for their help and encouragement.

REFERENCES

[1] W. L. Anderson and H. Diao, *Two-dimensional wavelet transform and application to holographic particle velocimetry*, Appl. Optics, 34 (1995), pp. 249–255.

[2] J.-P. Antoine, P. Carrette, R. Murenzi, and B. Piette, *Image analysis with two-dimensional continuous wavelet transform*, Signal Process., 31 (1993), pp. 241–272.

[3] D. Bernier and K. F. Taylor, *Wavelets from square integrable representations*, SIAM J. Math. Anal., 27 (1996), pp. 594–608.

[4] M. Duflo and C. C. Moore, *On the regular representation of a nonunimodular locally compact group*, J. Funct. Anal., 21 (1976), pp. 209–243.

[5] H. G. Feichtinger and K. H. Gröchenig, *Banach spaces related to integrable group representations and their atomic decompositions*, I, J. Funct. Anal., 86 (1989), pp. 307–340.

[6] A. Grossmann and J. Morlet, *Decomposition of Hardy functions into square integrable wavelets of constant shape*, SIAM J. Math. Anal., 15 (1984), pp. 723–736.

[7] A. Grossmann, J. Morlet, and T. Paul, *Transforms associated to square integrable group representations.* I. *General results*, J. Math. Phys., 26 (1985), pp. 2473–2479.

[8] C. E. Heil and D. F. Walnut, *Continuous and discrete wavelet transforms*, SIAM Rev., 31 (1989), pp. 628–666.
[9] S. Mallat, *Multifrequency channel decompositions of images and wavelet models*, IEEE Trans. Acoust. Speech Signal Process., 37 (1989), pp. 2091–2110.
[10] D. Mendlovic, I. Ouzieli, I. Kiryuschev, and E. Marom, *Two dimensional wavelet transform achieved by computer-generated multireference matched filter and Dammann grating*, Appl. Optics, 34 (1995), pp. 8213–8219.
[11] R. Murenzi, *Wavelet transforms associated to the n-dimensional Euclidean group with dilations: Signal in more than one dimension*, in Wavelets, Time-Frequency Methods and Phase Space (Proc. Marseille, December 1987), J.-M. Combs, A. Grossman, and P. Tchamitchian, eds., Springer-Verlag, Berlin, 1989, pp. 239–246.
[12] J. Phillips, *A note on square-integrable representations*, J. Funct. Anal., 20 (1975), pp. 83–92.
[13] S. Phuvan, *Optical implementation of N-wavelet coding for pattern classification*, Appl. Optics, 33 (1994), pp. 5294–5302.
[14] D. Roberge and Y. Sheng, *Optical wavelet matched filter*, Appl. Optics, 33 (1994), pp. 5287–5293.
[15] D. A. Russell and R. Ebert, *Optical realization of the wavelet transform for two-dimensional objects*, Appl. Optics, 32 (1993), pp. 6542–6546.
[16] D.-X. Wang, J.-W. Tai, and Y.-X. Zhang, *Two-dimensional optical wavelet transform in space domain and its performance analysis*, Appl. Optics, 33 (1994), pp. 5271–5274.

SIAM J. MATH. ANAL.
Vol. 32, No. 1, pp. 103–118

NONLINEAR ELLIPTIC PROBLEMS UNDER MIXED BOUNDARY VALUE CONDITIONS IN NONSMOOTH DOMAINS*

CARSTEN EBMEYER†

Abstract. Strongly nonlinear elliptic equations and systems are investigated under mixed boundary value conditions. It is supposed that the domain is a multidimensional polyhedral domain. Global regularity results of $|\nabla u|^\sigma$ ($\sigma \geq 1$) are proven, in particular, $W^{s,p}(\Omega)$-regularity ($s < \frac{1}{2}$) of $|\nabla u|$ and $|\nabla u|^2$.

Key words. mixed boundary value problem, piecewise smooth boundary, difference quotient, Nikolskii space.

AMS subject classifications. Primary, 35J55, 35J65; Secondary, 35J25.

PII. S0036141098349868

1. Introduction.

Let $\Omega \subset \mathbb{R}^n$ ($n \geq 2$) be a bounded polyhedral domain, let $\partial\Omega = \Gamma_D \cup \Gamma_N$, and let ν be the outward normal of $\partial\Omega$. We consider the mixed boundary value problem

$$\begin{aligned} -\sum_{i=1}^n \partial_i F_i(x, \nabla u) &= f(x) + \sum_{i=1}^n \partial_i f_i(x) && \text{in } \Omega, \\ u(x) &= 0 && \text{on } \Gamma_D, \\ -\sum_{i=1}^n F_i(x, \nabla u)\nu_i &= \sum_{i=1}^n f_i \nu_i && \text{on } \Gamma_N, \end{aligned} \tag{1.1}$$

where Γ_D is the Dirichlet and Γ_N the Neumann boundary.

We treat elliptic equations and systems. The aim of this paper is to investigate the regularity of $|\nabla u|^\sigma$ for $\sigma \geq 1$. We prove $W^{s,p}(\Omega)$-regularity results of $|\nabla u|^\sigma$ up to the boundary for $s = \frac{1}{2} - \varepsilon$.

It is known that solutions of mixed boundary value problems may have singularities on the boundary at points where the boundary condition changes or where the boundary is not smooth. In the case of linear elliptic equations various regularity results in Sobolev spaces of fractional order have been proven. In the case when the domain Ω is two-dimensional, detailed information about the behavior of u near singular points is known (see, eg., [1, 6, 10, 11]). In higher dimensions (i.e., $\Omega \subset \mathbb{R}^n$ and $n > 2$) there are still open questions even if the equation is linear. Three-dimensional problems are investigated in [2, 3, 15, 16, 20, 21]. In the case when the equation is nonlinear the literature is very rare. Quasi-linear equations and problems with p-structure are considered in [4, 5, 17, 22], where barrier functions are constructed. Asymptotic expansions of solutions are given in [7, 18].

In [8, 9] problem (1.1) is studied. A new technique is developed in order to treat strongly nonlinear problems. The regularity of a weak solution u is investigated. It is shown that $u \in W^{\frac{3}{2}-\varepsilon,2}(\Omega)$ and $\nabla u \in L^{3+\varepsilon}(\Omega)$.

*Received by the editors December 28, 1998; accepted for publication (in revised form) January 21, 2000; published electronically June 22, 2000.
http://www.siam.org/journals/sima/32-1/34986.html
†Mathematisches Seminar, Universität Bonn, Nussallee 15, D-53115 Bonn, Germany (ebmeyer@uni-bonn.de).

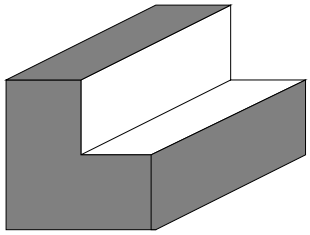


FIG. 1. *An admissible nonconvex domain where angle*$(\Gamma_N, \Gamma_D) \leq \pi$.

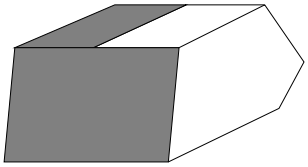


FIG. 2. *An admissible domain in the case when angle*$(\Gamma_N, \Gamma_D) = \pi$.

In this paper, we prove regularity results of $|\nabla u|^\sigma$. Therefore, we have refined and simplified the technique of [8, 9]. The method of proof is a difference quotient technique. Let us emphasize that we do not use Fourier series as in [8, 9]. Further, following the proof given below we could obtain all results of [8, 9] without using Fourier series.

This paper is organized as follows. In the next section we give the assumptions on the data and state the main results. In section 4 we investigate the regularity of certain difference quotients of ∇u. Section 5 contains the proofs of the main theorems.

2. Assumptions on the data and the main results. The domain $\Omega \subset \mathbb{R}^n$ ($n \geq 2$) is a bounded polyhedral domain. We assume that $N \geq 1$, $u : \Omega \to \mathbb{R}^N$, $f, f_i : \Omega \to \mathbb{R}^N$, and $F_i : \Omega \times \mathbb{R}^{nN} \to \mathbb{R}^N$ for $1 \leq i \leq n$.

In order to state our first theorem we give the assumptions on the functions F_i, f_i, and f in the case when $N = 1$. First of all, we suppose that

(i) $\partial\Omega = \bigcup_{1 \leq i \leq M} \overline{\Gamma^i}$, where each Γ^i is an open subset of a hyperplane, and $\partial\Gamma^i$ is polyhedral.

(ii) $\Gamma^i \cap \Gamma^j = \emptyset$ for $i \neq j$.

(iii) $\Gamma^i \subset \Gamma_D$ or $\Gamma^i \subset \Gamma_N$ for each $1 \leq i \leq M$.

(iv) $\bigcap_{i \in \Lambda} \partial\Gamma^i = \emptyset$ if $\Lambda \subset \{1, \ldots, M\}$ and $|\Lambda| > n$.

(v) angle$(\Gamma^i, \Gamma^j) \leq \pi$ if $\Gamma^i \subset \Gamma_D$, $\Gamma^j \subset \Gamma_N$, and $\partial\Gamma^i \cap \partial\Gamma^j \neq \emptyset$.

(vi) There is at most one pair of boundary manifolds Γ^i, Γ^j $(1 \leq i, j \leq M)$ satisfying $\Gamma^i \subset \Gamma_D$, $\Gamma^j \subset \Gamma_N$, $\partial\Gamma^i \cap \partial\Gamma^j \neq \emptyset$, and angle$(\Gamma^i, \Gamma^j) = \pi$.

Remark. (i) By angle(Γ^i, Γ^j) we denote the inner angle between a pair of boundary manifolds Γ^i, Γ^j satisfying $\partial\Gamma^i \cap \partial\Gamma^j \neq \emptyset$.

(ii) Let $\partial\Gamma^i \cap \partial\Gamma^j \neq \emptyset$. We assume that angle$(\Gamma^i, \Gamma^j) \leq \pi$, if $\Gamma^i \subset \Gamma_D$ and $\Gamma^j \subset \Gamma_N$. But if the boundary condition does not change (i.e., either $\Gamma^i, \Gamma^j \subset \Gamma_D$ or $\Gamma^i, \Gamma^j \subset \Gamma_N$) we admit $0 <$ angle$(\Gamma^i, \Gamma^j) < 2\pi$.

(iii) In Figure 1 and Figure 2 examples of admissible domains are given, where Γ_N and Γ_D are marked grey and white, respectively. Figure 1 shows a nonconvex domain. In Figure 2 a domain is given where the angle between a Dirichlet and a Neumann boundary manifold is equal to π.

Let $x \in \overline{\Omega}$, $r \in \mathbb{R}^n$, and $x = (x_1, \ldots, x_n)^T$. We suppose there is a C^2-function $F(x, r)$ such that $\frac{\partial}{\partial r_i} F(x, r) = F_i(x, r)$ for $1 \leq i \leq n$. We set $F_{x_i}(x, r) = \frac{\partial}{\partial x_i} F(x, r)$, $F_{i,x_k}(x, r) = \frac{\partial}{\partial x_k} F_i(x, r)$, and $F_{i,k}(x, r) = \frac{\partial}{\partial r_k} F_i(x, r)$ for $1 \leq i, k \leq n$. Furthermore, we suppose there are functions g_0, g_{x_i}, g_i, and g_{i,x_k} $(1 \leq i, k \leq n)$ (here the indices do not denote derivatives) such that

(H1) $c_0 + c_0'|r|^2 \leq F(x, r) \leq g_0(x) + c|r|^2$ for $g_0 \in L^\infty(\Omega)$ and $c_0' > 0$,

(H2) $|F_{x_i}(x, r)| \leq g_{x_i}(x) + c|r|^2$ for $g_{x_i} \in L^1(\Omega)$,

(H3) $|F_i(x,r)| \le g_i(x) + c|r|$ for $g_i \in L^2(\Omega)$,
(H4) $|F_{i,x_k}(x,r)| \le g_{i,x_k}(x) + c|r|$ for $g_{i,x_k} \in L^2(\Omega)$,
(H5) $|F_{i,k}(x,r)| \le c$,
(H6) there is a constant $k_0 > 0$ independent of x and r such that

$$k_0|\xi|^2 \le \sum_{i,k=1}^{n} F_{i,k}(x,r)\xi_i\xi_k \qquad \text{for all } \xi \in \mathbb{R}^n,$$

(H7) $f(x) \in L^2(\Omega)$ and $f_i(x) \in W^{1,2}(\Omega) \cap L^\infty(\Omega)$ for $1 \le i \le n$.

Remark. We assume there is a function $F(x,r)$ such that $F_i(x,r)$ is the partial derivative of F with respect to the ith component of r. Thus, our proof is restricted to the variational case.

Let $u : \Omega \to \mathbb{R}$ and $V = \{v \in W^{1,2}(\Omega) : v = 0 \text{ on } \Gamma_D\}$. It is well known that under the above hypotheses there exists a unique weak solution $u(x) \in W^{1,2}(\Omega)$ satisfying

$$\sum_{i=1}^{n} \int_\Omega F_i(x, \nabla u)\partial_i v = \int_\Omega fv - \sum_{i=1}^{n} \int_\Omega f_i \partial_i v \qquad \text{for all } v \in V. \tag{2.1}$$

We prove the following result for $N = 1$.

THEOREM 2.1. *Let $u : \Omega \to \mathbb{R}$ be a scalar function and $1 \le \sigma \le \frac{5}{2}$. Let the functions g_{x_i}, g_i, g_{i,x_k}, f, and f_k, given in* (H1)–(H7), *satisfy*

$$g_i \in L^{\frac{n}{1-\delta}}(\Omega), \text{ and } g_{x_i},\ g_{i,x_k},\ f,\ \partial_i f_k \in L^{\frac{2n}{3-\delta}}(\Omega) \tag{2.2}$$

for $1 \le i,k \le n$ and some small $\delta > 0$. Then it holds that

$$|\nabla u|^\sigma \in W^{s,p}(\Omega) \qquad \text{for all } s < \frac{1}{2} \text{ and } p = \frac{6}{2\sigma+1} + \varepsilon_0, \tag{2.3}$$

where $\varepsilon_0 = 0$ for $\sigma = 1$ and $\varepsilon_0 > 0$ for $\sigma > 1$.

Remark. (i) In the case when $n = 2$ and $\sigma > 1$, a better result than (2.3) can be proven; cf. Theorem 2.2 below.

(ii) In the proof of Theorem 2.1 we use the fact that $u \in C^{0,\alpha}(\Omega)$ for some $\alpha > 0$. In particular, if $\sigma > 1$ it holds that $\varepsilon_0 > 0$, where the value of ε_0 depends on α.

(iii) Clearly, it holds that $\nabla u \in W^{\frac{1}{2}-\varepsilon,2}(\Omega)$. This is proven in [8, 9]. Further, (2.3) implies that $|\nabla u|^2 \in W^{\frac{1}{2}-\varepsilon,p}(\Omega)$ for $p = \frac{6}{5}$ and $|\nabla u|^{\frac{5}{2}} \in W^{\frac{1}{2}-\varepsilon,1}(\Omega)$.

Our method of proof can also be applied to elliptic systems, i.e., $u : \Omega \to \mathbb{R}^N$ and $N > 1$. Let $x \in \Omega \subset \mathbb{R}^n$, $r \in \mathbb{R}^{nN}$, and let $F(x,r)$ satisfy the above growth conditions. Let us note that then the ellipticity condition reads as follows:

$$k_0|\xi|^2 \le \sum_{s,t=1}^{N} \sum_{i,k=1}^{n} F_{i,k}^{st}(x,r)\xi_i^s\xi_k^t \qquad \text{for all } \xi \in \mathbb{R}^{nN},$$

where $F_{i,k}^{st}(x,r) = \frac{\partial}{\partial r_k^t} F_i^s(x,r)$, and F_i^s is the sth component of F_i.

The proof of Theorem 2.1 employs the fact that u is Hölder continuous. But it is well known that this may not be satisfied if $N > 1$.

For solutions $u : \Omega \to \mathbb{R}^N$ $(N > 1)$ which are Hölder continuous, the proof of Theorem 2.1 provides (2.3). In the general case there holds the following result.

THEOREM 2.2. *Let $u : \Omega \to \mathbb{R}^N$, $N \ge 1$, and $1 \le \sigma < \frac{2n-1}{n-1}$. Then it holds that*

$$|\nabla u|^\sigma \in W^{s,p}(\Omega) \qquad \text{for all } s < \frac{1}{2} \text{ and } p = \frac{2}{1 + (\sigma-1)\frac{n-1}{n}}. \tag{2.4}$$

Remark. (i) We can also prove Theorems 2.1 and 2.2 for domains with a piecewise smooth boundary. Then we locally map the domain onto a polyhedron. Therefore we need that $\partial\Omega$ is $W^{2,\infty}$-piecewise. Details about this technique can be found in [8, 9]. Further, we can treat domains with a slit.

(ii) Let $n = 2$. In the case when the problem is linear it is known that $|u| \le cr^\alpha$ for some $\alpha \ge \frac{1}{2}$, where r denotes the distance to the corner point. This result holds for equations and systems as well; cf. the survey [12]. In particular, the function $u(r,\varphi) = r^{\frac{1}{2}} \sin\frac{1}{2}\varphi$ is a weak solution of

$$\Delta u(x,y) = 0 \text{ in } \mathbb{R}\times\mathbb{R}^+, \quad u(x,0) = 0 \text{ for } x \ge 0, \quad -\partial_y u(x,0) = 0 \text{ for } x < 0.$$

If $\Omega \subset \mathbb{R}\times\mathbb{R}^+$ is an appropriate bounded domain, u satisfies $|\nabla u|^\sigma \in W^{s,\frac{4}{\sigma+1}}(\Omega)$ for $s < \frac{1}{2}$, but $|\nabla u|^\sigma \notin W^{s,\frac{4}{\sigma+1}}(\Omega)$ for $s = \frac{1}{2}$. Thus, Theorem 2.2 is sharp for $n = 2$.

(iii) Let $n = 3$ and $N \ge 1$. If the domain has a slit and the boundary condition does not change, there holds $|u| \le cr^{\frac{1}{2}}$ where r is the distant to the edge; cf. [12]. Then, in general, $|\nabla u|^\sigma \in W^{s,\frac{4}{\sigma+1}}(\Omega)$ if and only if $s < \frac{1}{2}$. This implies that (2.3) and (2.4) are sharp for $\sigma = 1$.

(iv) If $n \ge 3$ and $\sigma > 1$, our results are possibly not sharp. Let us note that (2.3) yields a better result for Hölder continuous solutions than (2.4). Then the value of p depends on the size of the Hölder exponent.

3. Notations. Let $P \in \partial\Omega$, $R > 0$, $B_R(P) = \{x \in \mathbb{R}^n : |P - x| < R\}$, $B(R) = B_R(P) \cap \Omega$, and $B_R = B_R(P)$. We assume P and R are fixed such that P is the only vertex of $B_{3R} \cap \partial\Omega$ or that there is no vertex of $\partial\Omega$ in B_{3R}. Further, $B_{3R} \cap \partial\Omega$ is simply connected, and $\partial\Gamma^i \cap \partial\Gamma^j \ne \emptyset$ holds if $\Gamma^i \cap B_{3R} \ne \emptyset$ and $\Gamma^j \cap B_{3R} \ne \emptyset$.

For convenience, we set $B = B(2R)$ and $B_0 = B(R)$. In the case when $\Gamma^k \cap B_{3R} \ne \emptyset$ we often write Γ^k_* instead of $\Gamma^k \cap B_{3R}$.

The function τ_0 is a cut-off function satisfying $\tau_0 \equiv 1$ in $B_R(P)$, supp $\tau_0 = B_{2R}(P)$, and $|\nabla\tau_0| \le \frac{c}{R}$. By τ we denote the restriction of τ_0 onto $\overline{\Omega}$.

Let $\zeta^1, \dots, \zeta^n$ be a basis of $\mathbb{R}^n$. We assume $|\zeta^i| = 1$ for each $1 \le i \le n$ and $x + s\zeta^i \subset \overline{\Omega}$ for $x \in \overline{B}$ and $s \in (0,R)$. Below, we use the shift operator $E_i^\sigma x = x + \sigma\zeta^i$. We will write $E_i^\sigma f(x)$ instead of $f(E_i^\sigma x)$ and $E_i^\sigma f(x)g(x)$ instead of $(E_i^\sigma f(x))g(x)$. In what follows, we assume $h > 0$. Let us set

$$\Delta_i^h f(x) = E_i^h f(x) - f(x) \quad \text{and} \quad \Delta_i^{-h} f(x) = E_i^{-h} f(x) - f(x).$$

Moreover, we define the Sobolev spaces $W^{s,p}(\Omega)$ and the Nikolskii spaces $\mathcal{H}^{s,p}(\Omega)$; cf. [13]. Let m be an integer, $0 < \sigma < 1$, $s = m + \sigma$, $1 \le p < \infty$, $z \in \mathbb{R}^n$, and $\Omega_\eta = \{x \in \Omega : \operatorname{dist}(x,\partial\Omega) \ge \eta\}$. The spaces $W^{s,p}(\Omega)$ and $\mathcal{H}^{s,p}(\Omega)$ consist of all functions $f : \Omega \to \mathbb{R}$ for which the norms

$$\|f\|_{W^{s,p}(\Omega)} = \left(\|f\|^p_{W^{m,p}(\Omega)} + \sum_{|\alpha|=m} \int_\Omega \int_\Omega \frac{|\partial^\alpha f(x) - \partial^\alpha f(y)|^p}{|x-y|^{n+p\sigma}}\, dxdy \right)^{\frac{1}{p}}$$

and

$$\|f\|_{\mathcal{H}^{s,p}(\Omega)} = \left(\|f\|^p_{L^p(\Omega)} + \sum_{|\alpha|=m} \sup_{\substack{\eta>0 \\ 0<|z|<\eta}} \int_{\Omega_\eta} \frac{|\partial^\alpha f(x+z) - \partial^\alpha f(x)|^p}{|z|^{\sigma p}}\, dx \right)^{\frac{1}{p}}$$

are finite.

We write $\partial_i = \frac{\partial}{\partial x_i}$ and $\sum_i$ instead of $\sum_{i=1}^n$. Further, c denotes a constant which will be allowed to vary from equation to equation.

4. The basic estimate. In this section we prove $h^{-\frac{1}{2}+\varepsilon}\Delta_k^h \nabla u \in L^2(B_0)$ for all $\varepsilon > 0$; see Proposition 4.1 below. First, let us give some notations. Let ζ^1 be a basis vector. By definition, $x + h\zeta^1 \in \overline{\Omega}$ holds for all $x \in \overline{\Omega \cap B_{2R}}$. We define

$$\Omega_1^h = \{z \in B(3R) : z \neq x + h\zeta^1,\ x \in B(3R)\} \tag{4.1}$$

and

$$\Omega_1^{-h} = \{z \in B_{3R}(P) \setminus \Omega : z = x - h\zeta^1,\ x \in B(3R)\}. \tag{4.2}$$

Let $z_0 \in \partial\Omega \cap \partial\Omega_1^{-h}$, $0 < \lambda \leq h$, and $z_0 - \lambda\zeta^1 \in \Omega_1^{-h}$. For some function f_1 we define an even extension into Ω_1^{-h} by setting

$$f_1(z_0 - \lambda\zeta^1) = f_1(z_0 + \lambda\zeta^1). \tag{4.3}$$

Next, let f_2 be a function satisfying $f_2(z_0) = 0$ for all $z_0 \in \partial\Omega \cap \partial\Omega_1^{-h}$. We extend f_2 into Ω_1^{-h} by setting

$$f_2(z_0 - \lambda\zeta^1) = 0. \tag{4.4}$$

Let us note that this is an $W^{1,2}$-extension and, in fact, an $\mathcal{H}^{\frac{3}{2},2}$-extension.

Now we state the main result of this section.

PROPOSITION 4.1. *Let $\varepsilon > 0$. There is a basis $\zeta^1, \ldots, \zeta^n$ of $\mathbb{R}^n$ such that*

$$\sup_{0<h<R} \int_{B_0} \left| h^{-\frac{1}{2}+\varepsilon} \Delta_k^h \nabla u \right|^2 dx \leq c \qquad \text{for } 1 \leq k \leq n, \tag{4.5}$$

where the constant c depends only on R, ε, and the data.

In order to prove the proposition we proceed in several steps. Let $\Gamma_*^i = \Gamma^i \cap B_{3R}$. We set

$$\omega = \max\{\text{angle}(\Gamma_*^i, \Gamma_*^j) : \Gamma_*^i \subset \Gamma_N, \Gamma_*^j \subset \Gamma_D\}$$

if $\Gamma_N \cap B_{3R} \neq \emptyset$ and $\Gamma_D \cap B_{3R} \neq \emptyset$. Otherwise, we set $\omega = 0$. (By $\text{angle}(\Gamma_*^i, \Gamma_*^j)$ we denote the inner angle between Γ_*^i and Γ_*^j.) In what follows, we distinguish the cases when $\omega < \pi$ or when $\omega = \pi$.

LEMMA 4.2. *Let $\omega < \pi$. Let ζ^1 be parallel to $\Gamma_N \cap B_{3R}$ and angle$(\zeta^1, \Gamma_D \cap B_{3R}) \geq \alpha > 0$. Then there holds that*

$$\sup_{0<h<R} \int_{B_0} \left| h^{-\frac{1}{2}} \Delta_1^h \nabla u \right|^2 dx \leq c, \tag{4.6}$$

where the constant c depends only on R, α, and the data.

Proof. Let $0 < h < R$. We define extensions of the functions $\tau(\cdot)$ and $F(\cdot, r)$ into Ω_1^{-h} using (4.3), and we extend u using (4.4).

We choose the test function $v = \tau^2 h^{-1} \Delta_1^{-h} u \equiv \tau^2 h^{-1}(E_1^{-h}u - u)$ in (2.1). Notice that this is an admissible test function, for there holds $E_1^{-h}x \in \Omega_1^{-h}$ for $x \in \Gamma_D \cap B$, thus $E_1^{-h}u(x) = u(x) = 0$ on $\Gamma_D \cap B$.

Taking $v = \tau^2 h^{-1}\Delta_1^{-h}u$ in (2.1) we obtain

$$\begin{aligned}\sum_i \int_B F_i(x,\nabla u)\tau^2 h^{-1}\partial_i\Delta_1^{-h}u = &-\sum_i \int_B F_i(x,\nabla u)\partial_i\tau^2 h^{-1}\Delta_1^{-h}u\\ &+\int_B f\tau^2 h^{-1}\Delta_1^{-h}u\\ &-\sum_i \int_B f_i\partial_i\left(\tau^2 h^{-1}\Delta_1^{-h}u\right).\end{aligned} \tag{4.7}$$

Let $r \in \mathbb{R}^n$. The Taylor expansion of $F(x,r)$ and the ellipticity condition (H6) entail

$$\begin{aligned}F(x,r') - F(x,r) &= \sum_i (r'-r)_i F_i(x,r)\\ &\quad+\sum_{i,k}(r'-r)_i(r'-r)_k\int_0^1 (1-t)F_{i,k}(x,tr'+(1-t)r)\,dt\\ &\geq \sum_i (r'-r)_i F_i(x,r) + \frac{k_0}{2}|r'-r|^2.\end{aligned} \tag{4.8}$$

We put $r = \nabla u$ and $r' = E_1^{-h}\nabla u$. This yields

$$-\sum_i F_i(x,\nabla u)h^{-1}\Delta_1^{-h}\partial_i u \geq -h^{-1}[F(x,E_1^{-h}\nabla u) - F(x,\nabla u)] + \frac{k_0}{2h}\left|\Delta_1^{-h}\nabla u\right|^2.$$

In view of (4.7), we obtain

$$\begin{aligned}&\frac{k_0}{2}\int_B \tau^2 h^{-1}\left|\Delta_1^{-h}\nabla u\right|^2\\ &\leq \int_B \tau^2 h^{-1}\Delta_1^{-h}F(x,\nabla u)\\ &\quad+\int_B \tau^2 h^{-1}\left[F\left(x,E_1^{-h}\nabla u\right) - F\left(E_1^{-h}x,E_1^{-h}\nabla u\right)\right]\\ &\quad+\sum_i\int_B F_i(x,\nabla u)\partial_i\tau^2 h^{-1}\Delta_1^{-h}u\\ &\quad-\int_B f\tau^2 h^{-1}\Delta_1^{-h}u + \sum_i\int_B f_i\partial_i\left(\tau^2 h^{-1}\Delta_1^{-h}u\right)\\ &= J_1+\cdots+J_5.\end{aligned}$$

Now we estimate the integrals $J_1,\dots,J_5$. Let us define

$$B_h = \{x \in B(3R) : x = y + \lambda\zeta^1,\ y \in B,\ 0 \leq \lambda \leq h\}. \tag{4.9}$$

The Leibniz rule $\Delta_1^{-h}fg = \Delta_1^{-h}(fg) - E_1^{-h}f\Delta_1^{-h}g$ implies that

$$\begin{aligned}J_1 &= \int_{B_h}\tau^2 h^{-1}\Delta_1^{-h}F(x,\nabla u)\\ &= \int_{B_h} h^{-1}\Delta_1^{-h}\left(\tau^2 F(x,\nabla u)\right) - \int_{B_h} h^{-1}\Delta_1^{-h}\tau^2 E_1^{-h}F(x,\nabla u)\\ &= J_{11}+J_{12}.\end{aligned}$$

The extension (4.4) yields $\nabla u = 0$ in Ω_1^{-h}, where Ω_1^{-h} is introduced in (4.2). Due to (H1) it follows that

$$\begin{aligned} J_{11} &= h^{-1} \int_{\Omega_1^{-h}} \tau^2 F(x, \nabla u) \le h^{-1} \int_{\Omega_1^{-h}} |g_0| \\ &\le \|g_0\|_{L^\infty(\Omega_1^{-h})}\, h^{-1} |\Omega_1^{-h}| \le c. \end{aligned}$$

Using (H1) again and the fact that $\tau^2 \in W^{1,\infty}(\Omega)$ we get

$$|J_{12}| \le c \left(\left\| E_1^{-h} g_0 \right\|_{L^1(B_h)} + \left\| E_1^{-h} \nabla u \right\|^2_{L^2(B_h)} \right) \le c.$$

Let $\zeta^1 = (\zeta_1^1, \dots, \zeta_n^1)^T$. Hypothesis (H2) and the Taylor expansion yield

$$\begin{aligned} |J_2| &\le \int_B \tau^2 \sum_k \left|\zeta_k^1\right| \int_0^1 \left| F_{x_k} \left(x - (1-t) h \zeta^1, E_1^{-h} \nabla u \right) \right| \, dt dx \\ &\le c \left(\sum_k \sup_{0 \le t \le 1} \left\| g_{x_k}(x - th\zeta^1) \right\|_{L^1(B)} + \left\| E_1^{-h} \nabla u \right\|^2_{L^2(B)} \right) \\ &\le c. \end{aligned}$$

In view of (H3) and (H7), we obtain

$$|J_3| \le c \left(\sum_i \|g_i\|^2_{L^2(B)} + \|\nabla u\|^2_{L^2(B)} + \left\| h^{-1} \Delta_1^{-h} u \right\|^2_{L^2(B)} \right) \le c$$

and

$$|J_4| \le c \left(\|f\|^2_{L^2(B)} + \left\| h^{-1} \Delta_1^{-h} u \right\|^2_{L^2(B)} \right) \le c.$$

Further, let B_h be the set introduced in (4.9). We find

$$J_5 = \sum_i \int_{B_h} f_i \partial_i \tau^2 h^{-1} \Delta_1^{-h} u + \sum_i \int_{B_h} f_i \tau^2 h^{-1} \Delta_1^{-h} \partial_i u = J_{51} + J_{52}.$$

Hypothesis (H7) yields

$$|J_{51}| \le c \left(\|f_i\|^2_{L^2(B_h)} + \left\| h^{-1} \Delta_1^{-h} u \right\|^2_{L^2(B_h)} \right) \le c.$$

Applying the Leibniz rule we get

$$\begin{aligned} J_{52} &= \sum_i \int_{B_h} h^{-1} \Delta_1^{-h} (f_i \tau^2 \partial_i u) - \sum_i \int_{B_h} h^{-1} \Delta_1^{-h} (f_i \tau^2) E_1^{-h} \partial_i u \\ &= J_{53} + J_{54}. \end{aligned}$$

Let us note that $\partial_i u = 0$ in Ω_1^{-h}, thus,

$$J_{53} = h^{-1} \sum_i \int_{\Omega_1^{-h}} \tau^2 f_i \partial_i u = 0.$$

Moreover, noting that $\tau \in W^{1,\infty}(\Omega)$ we get

$$|J_{54}| \le c \left(\|f_i\|^2_{W^{1,2}(B_h)} + \|\nabla u\|^2_{L^2(B_h)} \right) \le c.$$

Altogether we obtain

$$\int_B \tau^2 \left| h^{-\frac{1}{2}} \Delta_1^{-h} \nabla u \right|^2 \leq c$$

and the constant c is independent of h. Notice that $\tau \equiv 1$ in B_0. Thus the assertion follows. □

LEMMA 4.3. *Let $\omega < \pi$. Let ζ^1 be parallel to $\Gamma_D \cap B_{3R}$ and angle$(\zeta^1, \Gamma_N \cap B_{3R}) \geq \alpha > 0$. Then there holds that*

$$\sup_{0<h<R} \int_{B_0} \left| h^{-\frac{1}{2}} \Delta_1^h \nabla u \right|^2 dx \leq c, \tag{4.10}$$

where the constant c depends only on R, α, and the data.

Proof. Let $0 < h < R$. Notice that

$$x + \lambda \zeta^1 \in \Gamma_D \qquad \text{for all } x \in \Gamma_D \cap B_{2R} \text{ and } 0 < \lambda < R.$$

Thus the function $v = \tau^2 h^{-1} \Delta_1^h u$ is an admissible test function in (2.1). Testing the equation we obtain

$$\begin{aligned} \sum_i \int_B F_i(x, \nabla u) \tau^2 h^{-1} \partial_i \Delta_1^h u = &- \sum_i \int_B F_i(x, \nabla u) \partial_i \tau^2 h^{-1} \Delta_1^h u \\ &+ \int_B f \tau^2 h^{-1} \Delta_1^h u - \sum_i \int_B f_i \partial_i \left(\tau^2 h^{-1} \Delta_1^h u \right). \end{aligned}$$

Next, let us put $r = \nabla u$ and $r' = E_1^h \nabla u$ in the Taylor expansion (4.8). This entails

$$- \sum_i F_i(x, \nabla u) h^{-1} \Delta_1^h \partial_i u \geq -h^{-1} [F(x, E_1^h \nabla u) - F(x, \nabla u)] + \frac{k_0}{2h} \left| \Delta_1^h \nabla u \right|^2.$$

Thus we get

$$\begin{aligned} \frac{k_0}{2} \int_B \tau^2 h^{-1} \left| \Delta_1^h \nabla u \right|^2 \leq &\int_B \tau^2 h^{-1} \Delta_1^h F(x, \nabla u) \\ &+ \int_B \tau^2 h^{-1} \left[F\left(x, E_1^h \nabla u\right) - F\left(E_1^h x, E_1^h \nabla u\right) \right] \\ &+ \sum_i \int_B F_i(x, \nabla u) \partial_i \tau^2 h^{-1} \Delta_1^h u \\ &- \int_B f \tau^2 h^{-1} \Delta_1^h u + \sum_i \int_B f_i \partial_i \left(\tau^2 h^{-1} \Delta_1^h u \right) \\ = &\, J_1 + \cdots + J_5. \end{aligned}$$

The Leibniz rule $f \Delta_1^h g = \Delta_1^h (fg) - \Delta_1^h f E_1^h g$ yields

$$\begin{aligned} J_1 &= \int_B h^{-1} \Delta_1^h \left(\tau^2 F(x, \nabla u) \right) - \int_B h^{-1} \Delta_1^h \tau^2 E_1^h F(x, \nabla u) \\ &= J_{11} + J_{12} \end{aligned}$$

and it holds that

$$J_{11} = -h^{-1} \int_{\Omega_1^h} \tau^2 F(x, \nabla u).$$

Further, let us consider J_5. Using the notations from above we find

$$\begin{aligned}|J_{53}| &\equiv \left|\sum_i \int_B h^{-1}\Delta_1^h(f_i\tau^2\partial_i u)\right| = \left|\sum_i h^{-1}\int_{\Omega_1^h} f_i\tau^2\partial_i u\right| \\ &\le \frac{c}{\delta h}|\Omega_1^h|\sum_i \|\tau f_i\|^2_{L^\infty(\Omega_1^h)} + \frac{\delta}{h}\int_{\Omega_1^h}\tau^2|\nabla u|^2 \\ &\le c - \frac{\delta}{c_0'}J_{11}.\end{aligned}$$

Choosing $\delta = \frac{c_0'}{2}$ and using (H1) we obtain

$$J_{11} + |J_{53}| = c - \frac{1}{2h}\int_{\Omega_1^h}\tau^2 F(x,\nabla u) \le c - \frac{1}{2h}\int_{\Omega_1^h} c_0 \le c.$$

Now, following the proof of Lemma 4.2, we are able to estimate the integrals $J_1,\dots,J_5$. We may conclude that

$$\int_B \tau^2 h^{-1}\left|\Delta_1^h\nabla u\right|^2 \le c.$$

This yields the assertion. ∎

Next we investigate the regularity in the case when $\omega = \pi$. Let $\Gamma_*^i \subset \Gamma_N$, $\Gamma_*^j \subset \Gamma_D$, and $\operatorname{angle}(\Gamma_*^i,\Gamma_*^j) = \pi$. To begin with, let us assume

$$(\Gamma_*^i \cup \Gamma_*^j) \subset \{x\in\mathbb{R}^n : x_n = 0\} \quad\text{and}\quad \Omega\cap B_{3R} \subset \{x\in\mathbb{R}^n : x_n > 0\}. \tag{4.11}$$

In general, (4.11) is not satisfied. Then we make a rotation of the domain. This will be discussed later.

Further, let $\zeta^1,\dots,\zeta^n$ be a suitable basis of $\mathbb{R}^n$ such that $\zeta^n = e_n$ and $\zeta^i\cdot e_n = 0$ for $1\le i\le n-1$. We proceed in several steps.

LEMMA 4.4. *Let $\omega = \pi$ and let (4.11) be satisfied. There are basis vectors $\zeta^1,\dots,\zeta^n$ of $\mathbb{R}^n$ such that $\zeta^n = e_n$, $\zeta^i\cdot e_n = 0$ for $1\le i\le n-1$, and*

$$\sup_{0<h<R}\int_{B_0}\left|h^{-\frac12}\Delta_j^h\nabla u\right|^2 dx \le c \qquad \text{for } 1\le j\le n-1. \tag{4.12}$$

Further, the constant c depends only on R and the data.

Proof. Let $0<h<R$. Notice that $\operatorname{angle}(\Gamma_*^i,\Gamma_*^j) = \pi$. Let $\Gamma_1 = \overline{\Gamma_*^i\cup\Gamma_*^j}$ and $\Gamma_2 = (\partial\Omega\cap B_{3R})\setminus\Gamma_1$. We can find basis vectors ζ^i $(1\le i\le n-1)$ parallel to $\{x\in\mathbb{R}^n : x_n = 0\}$ such that

(1) $x + s\zeta^i \in \Gamma_D$ for $x\in\Gamma_1\cap\Gamma_D$ and $0<s<R$.
(2) $x + s\zeta^i \in \overline{\Omega}$ for $x\in\Gamma_2$ and $0<s<R$.
(3) At least one of the following conditions holds:
  (i) ζ^i is parallel to $\partial\Omega\cap B_{3R}$.
  (ii) There is just one Γ^k such that $\Gamma^k\cap B_{3R}\ne\emptyset$ and $\operatorname{angle}(\zeta^i,\Gamma^k)\ge\alpha$.
(4) $\operatorname{Angle}(\zeta^i,\zeta^j)\ge\alpha>0$ for $1\le i<j\le n-1$.
(5) α depends only on the geometry of $\partial\Omega$.

In order to prove (4.12) we distinguish three cases.

Case 1: ζ^i satisfies (3)(i). Then the function $v = \tau^2 h^{-1}\Delta_i^h u$ is an admissible test function in (2.1).

Case 2: ζ^i satisfies (3)(ii) where $\Gamma^k \subset \Gamma_N$. Then $v = \tau^2 h^{-1} \Delta_i^h u$ is an admissible test function.

Case 3: ζ^i satisfies (3)(ii) where $\Gamma^k \subset \Gamma_D$. Then $v = \tau^2 h^{-1} \Delta_i^{-h} u$ is an admissible test function.

Proceeding as in the proof of either Lemma 4.2 or Lemma 4.3 yields the assertion. ∎

LEMMA 4.5. *Let $\omega = \pi$, let (4.11) be satisfied, and $\varepsilon > 0$. Then there is a constant c depending only on R, ε, and the data such that*

$$\sup_{0<h<R} \int_{B_0} \left| h^{-\frac{1}{2}+\varepsilon} \Delta_n^h \partial_j u \right|^2 dx \le c \qquad \text{for } 1 \le j \le n-1. \tag{4.13}$$

Proof. Let $0 < h < R$ and $1 \le j \le n-1$. Estimate (4.12) entails

$$h^{-\frac{1}{2}} \Delta_j^h (\tau^2 u) \in W^{1,2}(\Omega).$$

The imbedding $W^{1,2}(\Omega) \to \mathcal{H}^{1,2}(\Omega)$ implies that $h^{-\frac{1}{2}} \Delta_j^h (\tau^2 u) \in \mathcal{H}^{1,2}(\Omega)$, thus,

$$h^{-\frac{1}{2}-\varepsilon} \Delta_i^h \left(h^{-\frac{1}{2}} \Delta_j^h (\tau^2 u) \right) \in \mathcal{H}^{\frac{1}{2}-\varepsilon,2}(\Omega) \qquad \text{for } 1 \le i \le n$$

and

$$h^{-\frac{1}{2}+\varepsilon} \Delta_l^h \left(\bar{h}^{-\frac{1}{2}-\varepsilon} \Delta_i^{\bar{h}} \left(\bar{h}^{-\frac{1}{2}} \Delta_j^{\bar{h}} (\tau^2 u) \right) \right) \in L^2(\Omega) \qquad \text{for } 1 \le i,l \le n,\ 0 < h, \bar{h} < R.$$

Choosing $i = j$ and $l = n$, we obtain

$$\bar{h}^{-1-\varepsilon} \Delta_j^{\bar{h}} \Delta_j^{\bar{h}} \left(h^{-\frac{1}{2}+\varepsilon} \Delta_n^h (\tau^2 u) \right) \in L^2(\Omega).$$

It follows that $\partial_{\zeta^j} (h^{-\frac{1}{2}+\varepsilon} \Delta_n^h u) \in L^2(B_0)$ for $0 < h < R$ and each $\varepsilon > 0$. This holds for all $j \in \{1, \dots, n-1\}$. Due to the fact that $\zeta^j \cdot e_n = 0$, the assertion follows. ∎

LEMMA 4.6. *Let $\omega = \pi$ and let (4.11) be satisfied. Further, let $\zeta^1, \dots, \zeta^n$ be the basis chosen in Lemma 4.4. Then there holds that*

$$\sup_{0<h<R} \int_{B_0} \left| h^{-\frac{1}{2}} \Delta_j^h F_n(x, \nabla u) \right|^2 dx \le c \qquad \text{for } 1 \le j \le n-1, \tag{4.14}$$

where the constant c depends only on R and the data.

Proof. Let $0 < h < R$, $1 \le j \le n-1$, and $\zeta^j = (\zeta_1^j, \dots, \zeta_n^j)^T$. The Taylor expansion, (H4), and (H5) entail

$$\begin{aligned}
&\int_{B_0} \left| h^{-\frac{1}{2}} \Delta_j^h F_n(x, \nabla u) \right|^2 \\
&\le \int_{B_0} \sum_k \left| h^{\frac{1}{2}} \zeta_k^j \int_0^1 F_{n,x_k}(x + th\zeta^j, \nabla u)\ dt \right|^2 \\
&\quad + \int_{B_0} \left(\sum_k \left| \int_0^1 F_{n,k}(E_j^h x, t E_j^h \nabla u + (1-t)\nabla u)\ dt \right| \left| h^{-\frac{1}{2}} \Delta_j^h \partial_k u \right| \right)^2 \\
&\le ch \left(\sum_k \sup_{0 \le t \le 1} \left\| g_{n,x_k}(x + th\zeta^j) \right\|_{L^2(B_0)}^2 + \|\nabla u\|_{L^2(B_0)}^2 \right) \\
&\quad + c \left\| h^{-\frac{1}{2}} \Delta_j^h \nabla u \right\|_{L^2(B_0)}^2.
\end{aligned}$$

Note that $g_{n,x_k} \in L^2(B)$. Further, due to (4.12), the last term on the right-hand side is bounded. Thus the assertion follows. ∎

LEMMA 4.7. *Let $\omega = \pi$, let (4.11) be satisfied, and let $\varepsilon > 0$. Then there holds that*

$$\sup_{0<h<R} \int_{B_0} \left| h^{-\frac{1}{2}+\varepsilon} \Delta_n^h F_n(x, \nabla u) \right|^2 dx \le c, \tag{4.15}$$

where the constant c depends only on R, ε, and the data.

Proof. Let $0 < h < R$ and $\varepsilon > 0$. Below we prove

$$\partial_j \left(h^{-\frac{1}{2}+\varepsilon} \Delta_n^h F_n(x, \nabla u) \right) \in W^{-1,2}(B_0) \qquad \text{for } 1 \le j \le n. \tag{4.16}$$

Further, let us note that $h^{-\frac{1}{2}+\varepsilon} \Delta_n^h F_n(x, \nabla u) \in W^{-1,2}(B_0)$. Now we apply the well-known estimate (cf. [19])

$$\|w\|_{L^2(\Omega)} \le c \left(\sum_i |\partial_i w|_{W^{-1,2}(\Omega)} + |w|_{W^{-1,2}(\Omega)} \right)$$

and obtain $h^{-\frac{1}{2}+\varepsilon} \Delta_n^h F_n(x, \nabla u) \in L^2(B_0)$ for $0 < h < R$. Thus the assertion follows.

In order to show (4.16), we distinguish two cases.

Case 1: *Let* $1 \le j \le n-1$. Estimate (4.14) yields $h^{-\frac{1}{2}} \Delta_j^h \left(\tau^2 F_n(x, \nabla u) \right) \in L^2(\Omega)$. It follows that

$$h^{-\frac{1}{2}+\varepsilon} \Delta_n^h \left(\bar{h}^{-\frac{1}{2}-\varepsilon} \Delta_j^{\bar{h}} \left(\bar{h}^{-\frac{1}{2}} \Delta_j^{\bar{h}} (\tau^2 F_n(x, \nabla u)) \right) \right) \in W^{-1,2}(\Omega)$$

for $0 < \bar{h} < R$, thus,

$$\partial_{\zeta^j} \left(h^{-\frac{1}{2}+\varepsilon} \Delta_n^h F_n(x, \nabla u) \right) \in W^{-1,2}(B_0).$$

Due to the fact that $\zeta^j \cdot e_n = 0$ for each $j \in \{1, \dots, n-1\}$ we get

$$\partial_j \left(h^{-\frac{1}{2}+\varepsilon} \Delta_n^h F_n(x, \nabla u) \right) \in W^{-1,2}(B_0) \qquad \text{for } 1 \le j \le n-1. \tag{4.17}$$

Case 2: *Let* $j = n$. Equation (1.1) yields

$$\begin{aligned} &h^{-\frac{1}{2}+\varepsilon} \Delta_n^h \left(\partial_n F_n(x, \nabla u) \right) \\ &= h^{-\frac{1}{2}+\varepsilon} \Delta_n^h \left(- \sum_{k=1}^{n-1} \partial_k F_k(x, \nabla u) - f(x) - \sum_i \partial_i f_i(x) \right). \end{aligned} \tag{4.18}$$

As in the proof of Lemma 4.6 it follows that

$$h^{-\frac{1}{2}} \Delta_k^h (\tau^2 F_k(x, \nabla u)) \in L^2(\Omega) \qquad \text{for } 1 \le k \le n-1.$$

This implies that

$$h^{-\frac{1}{2}+\varepsilon} \Delta_n^h \left(\bar{h}^{-1-\varepsilon} \Delta_k^{\bar{h}} \Delta_k^{\bar{h}} \left(\tau^2 F_k(x, \nabla u) \right) \right) \in W^{-1,2}(\Omega) \qquad \text{for } 1 \le k \le n-1,$$

thus,

$$h^{-\frac{1}{2}+\varepsilon} \Delta_n^h \partial_k F_k(x, \nabla u) \in W^{-1,2}(B_0) \qquad \text{for } 1 \le k \le n-1.$$

Hence, the right-hand side of (4.18) is a $W^{-1,2}(B_0)$-function. This entails

$$\partial_n \left(h^{-\frac{1}{2}+\varepsilon} \Delta_n^h F_n(x, \nabla u) \right) \in W^{-1,2}(B_0). \tag{4.19}$$

Altogether, (4.17) and (4.19) provide (4.16). This yields the assertion. □

LEMMA 4.8. *Let $\omega = \pi$, let (4.11) be satisfied, and let $\varepsilon > 0$. Then there is a constant c depending only on R, ε, and the data such that*

$$\sup_{0<h<R} \int_{B_0} \left| h^{-\frac{1}{2}+\varepsilon} \Delta_n^h \nabla u \right|^2 dx \leq c. \tag{4.20}$$

Proof. The Taylor expansion yields

$$\begin{aligned} J_1 &= h^{-\frac{1}{2}+\varepsilon} \Delta_n^h \partial_n u \int_0^1 F_{n,n}(x, tE_n^h \nabla u + (1-t)\nabla u)\, dt \\ &= h^{-\frac{1}{2}+\varepsilon} \Delta_n^h F_n(x, \nabla u) - h^{\frac{1}{2}+\varepsilon} \int_0^1 F_{n,x_n}(x + the_n, E_n^h \nabla u)\, dt \\ &\quad - \sum_{k=1}^{n-1} h^{-\frac{1}{2}+\varepsilon} \Delta_n^h \partial_k u \int_0^1 F_{n,k}(x, tE_n^h \nabla u + (1-t)\nabla u)\, dt \\ &= J_2 + J_3 + J_4. \end{aligned} \tag{4.21}$$

Estimate (4.15) entails

$$\sup_{0<h<R} \int_{B_0} |J_2|^2 = \sup_{0<h<R} \int_{B_0} \left| h^{-\frac{1}{2}+\varepsilon} \Delta_n^h F_n(x, \nabla u) \right|^2 \leq c.$$

Hypothesis (H4) implies that

$$\begin{aligned} &\sup_{0<h<R} \int_{B_0} |J_3|^2 \\ &\leq \sup_{0<h<R} ch^{1+2\varepsilon} \left(\sup_{0\leq t\leq 1} \|g_{n,x_n}(x + the_n)\|^2_{L^2(B_0)} + c \left\| E_n^h \nabla u \right\|^2_{L^2(B_0)} \right) \leq c. \end{aligned}$$

Using (4.13) and (H5) we get

$$\sup_{0<h<R} \int_{B_0} |J_4|^2 \leq \sup_{0<h<R} c \sum_{k=1}^{n-1} \int_{B_0} \left| h^{-\frac{1}{2}+\varepsilon} \Delta_n^h \partial_k u \right|^2 \leq c.$$

Thus from (4.21) we may conclude that

$$\sup_{0<h<R} \int_{B_0} |J_1|^2 = \sup_{0<h<R} \int_{B_0} \left| h^{-\frac{1}{2}+\varepsilon} \Delta_n^h \partial_n u F_{nn} \right|^2 \leq c,$$

where $F_{nn} = \int_0^1 F_{n,n}(x, tE_n^h \nabla u + (1-t)\nabla u)\, dt$. In view of the ellipticity condition (H6) we have

$$\int_{B_0} |J_1|^2 \geq k_0^2 \int_{B_0} \left| h^{-\frac{1}{2}+\varepsilon} \Delta_n^h \partial_n u \right|^2.$$

It follows that

$$\sup_{0<h<R}\int_{B_0}\left|h^{-\frac{1}{2}+\varepsilon}\Delta_n^h\partial_n u\right|^2\le c. \tag{4.22}$$

Altogether, (4.13) and (4.22) entail the assertion. □

Proof of Proposition 4.1. If $\Gamma_*^j\subset\Gamma_D$, $\Gamma_*^k\subset\Gamma_N$, and $\operatorname{angle}(\Gamma_*^j,\Gamma_*^k)=\pi$ we set $\Gamma_1=\overline{\Gamma_*^j\cup\Gamma_*^k}$ and $\Gamma_2=(\partial\Omega\cap B_{3R})\setminus\Gamma_1$. We choose appropriate basis vectors ζ^i $(1\le i\le n)$ such that $|\zeta^i|=1$ and there hold

(1) $\operatorname{angle}(\zeta^i,\zeta^j)\ge\alpha>0$ for $1\le i<j\le n$.

(2) $x+s\zeta^i\in\overline{\Omega}$ for $x\in\overline{\Omega}\cap B_{2R}$ and $0<s<R$.

(3) If $\Gamma_*^k\subset\Gamma_D$ and ζ^i is parallel to Γ_*^k, then $y+s\zeta^i\in\Gamma_*^k$ for $y\in\Gamma_*^k\cap B_{2R}$ and $0<s<R$.

(4) If $\Gamma^k\cap B_{3R}\ne\emptyset$ and ζ^i is not parallel to Γ^k, then $\operatorname{angle}(\zeta^i,\Gamma^k)\ge\alpha>0$.

(5) If $\operatorname{angle}(\zeta^i,\Gamma_*^k)\ge\alpha>0$, then $x-s\zeta^i\notin\overline{\Omega}$ for all $x\in\Gamma_*^k$ and $0<s<R$.

(6) There holds at least one of the following conditions:

(i) ζ^i is parallel to $\partial\Omega\cap B_{3R}$,

(ii) $\operatorname{angle}(\zeta^i,\Gamma_D\cap B_{3R})\ge\alpha$ and either $\Gamma_N\cap B_{3R}=\emptyset$ or ζ^i is parallel to $\Gamma_N\cap B_{3R}$,

(iii) $\operatorname{angle}(\zeta^i,\Gamma_N\cap B_{3R})\ge\alpha$ and either $\Gamma_D\cap B_{3R}=\emptyset$ or ζ^i is parallel to $\Gamma_D\cap B_{3R}$,

(iv) $\operatorname{angle}(\zeta^i,\Gamma_1\cap B_{3R})\ge\alpha$ and either $\Gamma_2\cap B_{3R}=\emptyset$ or ζ^i is parallel to $\Gamma_2\cap B_{3R}$,

(v) $\operatorname{angle}(\zeta^i,\Gamma_*^k\cap B_{3R})\ge\alpha$ for just one Γ_*^k and ζ^i is parallel to $\Gamma_1\cap B_{3R}$.

(7) α depends only on the geometry of $\partial\Omega$.

Clearly we can find such a basis satisfying (1)–(7). Let us note that, due to the conditions (1)–(7), we can proceed as in the above proofs. Therefore, we must extend u across Γ_D in some direction ζ^i. Due to (6), however, we need no extension of u across Γ_N.

Following the proofs of Lemmas 4.2, 4.3, 4.4, and 4.8 yields the assertion in the case when (i) $\omega<\pi$ or when (ii) $\omega=\pi$ and (4.11) is satisfied.

Next, let us consider the case when $\omega=\pi$ and (4.11) is not satisfied. Then there are Γ_*^i, Γ_*^j such that $\Gamma_*^i\subset\Gamma_N$, $\Gamma_*^j\subset\Gamma_D$, and $\operatorname{angle}(\Gamma_*^i,\Gamma_*^j)=\pi$. Let $\zeta^1,\dots,\zeta^{n-1}$ be tangential to $\Gamma_*^i\cup\Gamma_*^j$ and ζ^n be parallel to $\Gamma_2\cap B_{3R}$. For simplicity, we asume ζ^n is the inner normal of $\Gamma_*^i\cup\Gamma_*^j$. We rotate the domain Ω such that ζ^n is mapped onto e_n. Let A be the matrix describing the rotation. Let $\hat{x}=Ax$, $\hat{u}(\hat{x})=u(A^{-1}\hat{x})$, etc., and $\widetilde{\partial}_i=\sum_k a_{ik}\partial_k$. Then there holds the equation

$$-\sum_{i=1}^n\widetilde{\partial}_i\hat{F}_i(\hat{x},\widetilde{\nabla}\hat{u})=\hat{f}(\hat{x})+\sum_{i=1}^n\widetilde{\partial}_i\hat{f}_i(\hat{x})\qquad\text{in }\hat{\Omega}.$$

Let us note that $\widetilde{\partial}_1,\dots,\widetilde{\partial}_{n-1}$ are derivatives in directions tangential to the boundary and $\widetilde{\partial}_n$ is the normal derivative. Thus (4.12) yields

$$\sup_{0<h<R}\int_{\hat{B}_0}\left|h^{-\frac{1}{2}+\varepsilon}\widetilde{\Delta}_j^h\widetilde{\nabla}\hat{u}\right|^2\le c \tag{4.23}$$

for $1\le j\le n-1$ (and $\varepsilon=0$). Further, following the proof of Lemma 4.8 we obtain (4.23) for $j=n$ and $\varepsilon>0$. Due to

$$\left|\widetilde{\Delta}_j^h\widetilde{\nabla}\hat{u}\right|^2=\left|A\widetilde{\Delta}_j^h\nabla\hat{u}\right|^2=\left|\widetilde{\Delta}_j^h\nabla\hat{u}\right|^2$$

and the substitution rule for integrals we get

$$\sup_{0<h<R}\int_{B_0}\left|h^{-\frac{1}{2}+\varepsilon}\Delta_j^h\nabla u\right|^2\le c\qquad\text{for } 1\le j\le n,$$

where the constant c depends on R, $\varepsilon>0$, and the data. □

5. The proofs of Theorems 2.1 and 2.2. In this section we prove the main results.

Let us note that $f\in L^q(\Omega)$ and $f_i\in L^{2q}(\Omega)$ for some $q>\frac{n}{2}$ and $1\le i\le n$. Thus it holds that $u\in C^{0,\alpha}\left(\overline{\Omega}\right)$ for some $\alpha>0$. This can be proven as in [14], where quasi-linear problems are treated.

From the Hölder continuity of u and the assumption (2.2) it follows that

$$\nabla u\in L^{3+\varepsilon}(\Omega)\qquad\text{for some }\varepsilon>0; \tag{5.1}$$

cf. [9]. We use this result in order to prove Theorem 2.1.

Proof of Theorem 2.1. Let $r\in\mathbb{R}^n$. The Taylor expansion of $G(r)=|r|^\sigma$ entails

$$G(r')-G(r)=\sum_i(r'-r)_i\int_0^1 G_i(tr'+(1-t)r)\,dt,$$

where $G_i(r)=\frac{\partial}{\partial r_i}G(r)=\sigma|r|^{\sigma-2}r_i$. Let $j\in\{1,\dots,n\}$ be fixed. Setting $r=\nabla u$ and $r'=E_j^hr$ we get

$$\Delta_j^h|\nabla u|^\sigma=\sum_i\Delta_j^h\partial_iu\int_0^1 G_i(tr'+(1-t)r)\,dt.$$

Let $\varepsilon>0$ be small. We find

$$\left|h^{-\frac{1}{2}+\varepsilon}\Delta_j^h|\nabla u|^\sigma\right|\le\sigma\left|h^{-\frac{1}{2}+\varepsilon}\Delta_j^h\nabla u\right|\int_0^1\left|tE_j^h\nabla u+(1-t)\nabla u\right|^{\sigma-1}dt. \tag{5.2}$$

Let $\sigma>1$. Proposition 4.1 yields $h^{-\frac{1}{2}+\varepsilon}\Delta_j^h\nabla u\in L^2(B_0)$. Further, (5.1) implies that $|tE_j^h\nabla u+(1-t)\nabla u|^{\sigma-1}\in L^s(\Omega)$ for $s=\frac{3+\varepsilon}{\sigma-1}$. It follows that

$$\left|h^{-\frac{1}{2}+\varepsilon}\Delta_j^h\nabla u\right|\int_0^1\left|tE_j^h\nabla u+(1-t)\nabla u\right|^{\sigma-1}dt\in L^q(B_0) \tag{5.3}$$

for $q=\frac{2s}{2+s}=\frac{6+\delta}{2\sigma+1}$ and some $\delta>0$. Let us note that $\frac{6+\delta}{2\sigma+1}>1$ holds for $1<\sigma\le\frac{5}{2}$. Thus, in view of (5.2) and (5.3), we have

$$\sup_{0<h<R}\int_{B_0}\left|h^{-\frac{1}{2}+\varepsilon}\Delta_j^h|\nabla u|^\sigma\right|^{\frac{6+\delta}{2\sigma+1}}\le c\qquad\text{for } 1\le j\le n\text{ and }\sigma>1. \tag{5.4}$$

Next, estimate (5.2) and Proposition 4.1 entail

$$\sup_{0<h<R}\int_{B_0}\left|h^{-\frac{1}{2}+\varepsilon}\Delta_j^h|\nabla u|\right|^2\le c\qquad\text{for } 1\le j\le n. \tag{5.5}$$

Let $p=\frac{6+\delta}{2\sigma+1}$ for $\sigma>1$ and $p=2$ for $\sigma=1$. Then (5.4) and (5.5) imply (for $\sigma>1$ and $\sigma=1$, respectively) that

$$\sup_{\substack{\eta>0\\0<|z|<\eta}}\int_{B_\eta}\left|\frac{|\nabla u(x+z)|^\sigma-|\nabla u(x)|^\sigma}{|z|^{\frac{1}{2}-\varepsilon}}\right|^p dx\le c\qquad\text{for } 1\le\sigma\le\frac{5}{2}, \tag{5.6}$$

where $B_\eta = \{x \in B_0 : \operatorname{dist}(x, \partial\Omega) \geq \eta\}$. Further, the constant c depends only on ε, the data, and R.

We can find finite sets of numbers $\{R_1, \dots, R_k\}$ and of points $\{P_1, \dots, P_k\}$ depending only on the geometry of $\partial\Omega$ such that $\overline{\Omega} \subset \bigcup_{i=1}^k B_{R_i}(P_i)$, and each ball $B_{R_i}(P_i)$ satisfies: If $B_{R_i}(P_i) \cap \partial\Omega \neq \emptyset$, then $B_{R_i}(P_i) \cap \partial\Omega$ is simply connected, and $\partial\Gamma^i \cap \partial\Gamma^j \neq \emptyset$ holds if $\Gamma^i \cap B_{R_i}(P_i) \neq \emptyset$ and $\Gamma^j \cap B_{R_i}(P_i) \neq \emptyset$. Further, P_i is the only vertex of $B_{R_i}(P_i) \cap \partial\Omega$ or there is no vertex of $\partial\Omega$ in $B_{R_i}(P_i)$. Thus it follows that

$$|\nabla u|^\sigma \in \mathcal{H}^{\frac{1}{2}-\varepsilon,p}(\Omega).$$

The imbedding theorem of Nikolskii spaces into Sobolev spaces (cf. [13])

$$\mathcal{H}^{s,p}(\Omega) \to W^{s-\varepsilon,p}(\Omega) \qquad \text{for all } \varepsilon > 0$$

entails $|\nabla u|^\sigma \in W^{\frac{1}{2}-2\varepsilon,p}(\Omega)$ for all $\varepsilon > 0$. This yields the assertion. □

Proof of Theorem 2.2. We proceed as in the proof of Proposition 4.1. Let us remark that now, $u, F_i, f, f_i \in \mathbb{R}^N$ and $\nabla u \in \mathbb{R}^{nN}$ for some $N > 1$. The changes to be made in the proof are obvious. Hence Proposition 4.1 yields

$$\sup_{\substack{\eta>0 \\ 0<|z|<\eta}} \int_{B_\eta} \left| \frac{\nabla u(x+z) - \nabla u(x)}{|z|^{\frac{1}{2}-\varepsilon}} \right|^2 dx \leq c,$$

where $B_\eta = \{x \in B_0 : \operatorname{dist}(x, \partial\Omega) \geq \eta\}$. Applying the same cover argument as in the proof of Theorem 2.1 we get

$$\nabla u \in \mathcal{H}^{\frac{1}{2}-\varepsilon,2}(\Omega; \mathbb{R}^{nN}). \tag{5.7}$$

The imbedding theorem of Nikolskii spaces into Sobolev spaces yields

$$\nabla u \in W^{\frac{1}{2}-\varepsilon,2}(\Omega; \mathbb{R}^{nN}) \qquad \text{for all } \varepsilon > 0. \tag{5.8}$$

Let $\sigma > 1$. From the Sobolev imbedding theorem it follows that

$$|\nabla u|^{\sigma-1} \in L^s(\Omega) \qquad \text{for } 1 \leq s < \frac{2n}{(\sigma-1)(n-1)}. \tag{5.9}$$

Let $1 \leq j \leq n$. Due to (5.7) and (5.9) it holds that

$$\left| h^{-\frac{1}{2}+\varepsilon} \Delta_j^h \nabla u \right| \int_0^1 \left| t E_j^h \nabla u + (1-t)\nabla u \right|^{\sigma-1} dt \in L^q(\Omega) \quad \text{for } q = \frac{2s}{2+s}. \tag{5.10}$$

Thus, (5.10) holds for all $q < p$, where $p = \frac{2}{1+(\sigma-1)\frac{n-1}{n}}$. Using estimate (5.2) we get

$$h^{-\frac{1}{2}+\varepsilon} \Delta_j^h |\nabla u|^\sigma \in L^q(\Omega) \qquad \text{for } 1 \leq q < p \text{ and } 1 \leq j \leq n. \tag{5.11}$$

This implies that $|\nabla u|^\sigma \in \mathcal{H}^{\frac{1}{2}-\varepsilon,q}(\Omega)$ for $1 \leq q < p$, thus,

$$|\nabla u|^\sigma \in \mathcal{H}^{\frac{1}{2}-2\varepsilon,p}(\Omega) \qquad \text{for all } \varepsilon > 0 \text{ and } \sigma > 1. \tag{5.12}$$

Next, in the case when $\sigma = 1$, estimate (5.2) and Proposition 4.1 yield

$$|\nabla u| \in \mathcal{H}^{\frac{1}{2}-\varepsilon,2}(\Omega) \qquad \text{for all } \varepsilon > 0. \tag{5.13}$$

From (5.12), (5.13), and the imbedding theorem of Nikolskii spaces into Sobolev spaces the assertion follows. □

REFERENCES

[1] J. Banasiak and G. F. Roach, *On mixed boundary value problems of Dirichlet oblique-derivative type in plane domains with piecewise differentiable boundary*, J. Differential Equations, 79 (1989), pp. 111–131.

[2] J. Banasiak, *On asymptotics of solutions of elliptic mixed boundary value problems of second-order in domains with vanishing edges*, SIAM J. Math. Anal., 23 (1992), pp. 1117–1124.

[3] J. Banasiak, *A counterexample in the theory of mixed boundary value problems for elliptic equations in non-smooth domains*, Demonstratio Math., 26 (1993), pp. 327–335.

[4] M. V. Borsuk, *Estimates of solutions to the Dirichlet problem for second-order nondivergent elliptic equations in a neighborhood of a conic point of the boundary*, Differential Equations, 30 (1994), pp. 94–99.

[5] M. V. Borsuk and V. A. Kondrat'ev, *Properties, near a corner point, of a solution of a Dirichlet problem for a second-order quasilinear elliptic equation*, Differential Equations, 24 (1988), pp. 1185–1190.

[6] M. Dauge, *Elliptic Boundary Value Problems on Corner Domains*, Lecture Notes in Math. 1341, Springer-Verlag, Berlin, 1988.

[7] M. Dobrowolski, *On quasilinear elliptic equations in domains with conical boundary points*, J. Reine Angew. Math., 394 (1989), pp. 186–195.

[8] C. Ebmeyer, *Mixed boundary value problems for nonlinear elliptic systems in n-dimensional Lipschitzian domains*, Z. Anal. Anwendungen, 18 (1999), pp. 539–555.

[9] C. Ebmeyer and J. Frehse, *Mixed boundary value problems for nonlinear elliptic equations in multidimensional non-smooth domains*, Math. Nachr., 203 (1999), pp. 47–74.

[10] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Pitman Advanced Publishing Program, Boston, London, Melbourne, 1985.

[11] V. A. Kondrat'ev, *Boundary value problems for elliptic equations in domains with conical and angular points*, Trans. Moscow Math. Soc., 16 (1967), pp. 227–313.

[12] V. A. Kozlov and V. G. Maz'ya, *Singularities in solutions to mathematical physics problems in non-smooth domains*, in Partial Differential Equations and Functional Analysis, Birkhäuser, Bosten, 1996, pp. 174–206.

[13] A. Kufner, O. John, and S. Fučik, *Function Spaces*, Academia, Prague, 1977.

[14] O. A. Ladyzhenskaya and N. N. Uraltseva, *Linear and Quasilinear Elliptic Equations*, Academic Press, New York, London, 1968.

[15] V. G. Maz'ya and J. Rossmann, *Über die Asymptotik der Lösungen elliptischer Randwertaufgaben in der Umgebung von Kanten*, Math. Nachr., 138 (1988), pp. 27–53.

[16] V. G. Maz'ya and J. Rossmann, *On the behaviour of solutions to the Dirichlet problem for second order elliptic equations near edges and polyhedral vertices with critical angles*, Z. Anal. Anwendungen, 13 (1994), pp. 19–47.

[17] E. Miersemann, *Zur gemischten Randwertaufgabe für die Minimalflächengleichung*, Math. Nachr., 115 (1984), pp. 125–136.

[18] E. Miersemann, *Asymptotic expansion of solutions of the Dirichlet problem for quasilinear elliptic equations of second order near a conical point*, Math. Nachr., 135 (1988), pp. 239–274.

[19] R. Carroll, G. Duff, J. Friberg, J. Gobert, P. Grisvard, J. Necas, R. Seeley, *Equations aux Dérivées Partielles*, Seminaire de Mathematiques Superieures 19, Presses de l'Université de Montréal, Montreal, 1966.

[20] T. von Petersdorff and E. P. Stephan, *Decompositions in edge and corner singularities for the solution of the Dirichlet problem of the Laplacian in a polyhedron*, Math. Nachr., 149 (1990), pp. 71–104.

[21] H. Reisman, *Second order elliptic boundary value problems in a domain with edges*, Comm. Partial Differential Equations, 6 (1981), pp. 1023-1042.

[22] P. Tolksdorf, *Regularity for a more general class of quasilinear elliptic equations*, J. Differential Equations, 51 (1984), pp. 126-150.

SIAM J. MATH. ANAL.
Vol. 32, No. 1, pp. 119–141

NONSYMMETRIC LORENZ ATTRACTORS FROM A HOMOCLINIC BIFURCATION*

CLARK ROBINSON†

Abstract. We consider a bifurcation of a flow in three dimensions from a double homoclinic connection to a fixed point satisfying a resonance condition between the eigenvalues. For correctly chosen parameters in the unfolding, we prove that there is a transitive attractor of Lorenz type. In particular we show the existence of a bifurcation to an attractor of Lorenz type which is semiorientable, i.e., orientable on one half and nonorientable on the other half. We do not assume any symmetry condition, so we need to discuss nonsymmetric one-dimensional Poincaré maps with one discontinuity and absolute value of the derivative always greater than one. We also apply these results to a specific set of degree four polynomial differential equations. The results do not apply to the actual Lorenz equations because they do not have enough parameters to adjust to make them satisfy the hypothesis.

Key words. attractors, Lorenz, homoclinic bifurcation

AMS subject classifications. 34C35, 58F13

PII. S0036141098343598

1. Introduction.

In previous papers, [12] and [13], we proved that there is a bifurcation for differential equations in three dimensions with a symmetry from a double homoclinic connection for a fixed point to an attractor of Lorenz type. This attractor could be either untwisted or twisted on both sides. In this paper we consider the situation without a symmetry: in particular, we show that there can be a bifurcation from a double homoclinic connection to an attractor which is twisted on one side but untwisted on the other side. We give basic assumptions which are sufficient for this to take place. We also verify that specific polynomial differential equations in three dimensions can realize this bifurcation.

The mathematical results of this paper give a framework for analyzing the computational and experimental output exhibiting an attractor of Lorenz type. The analysis of the one-dimensional maps given in section 2 shows what type of structures can occur for attractors of flows in three dimensions even in cases not covered by the bifurcation theorem given in this paper. In particular, one-dimensional maps for attractors can occur that are either transitive on a single interval or transitive on multiple intervals. Also, typical trajectories outside the attractor can approach the attractor in different ways depending on the existence or lack of a trapping region. The codimension three bifurcation itself can occur as the unfolding of a degenerate singularity in systems for which there are many parameters which can be adjusted. Finally, the main theorem shows that attractors which are twisted on one or both sides can occur for actual polynomial differential equations. These twists relate to the type of knotted orbits which occur on the attractor. For an introduction to knots and templates for three-dimensional flows see [5] or [17].

In this paper, a transversality assumption and the dominance of the strong stable eigenvalue are used with standard stable manifold theory to reduce the problem to a

*Received by the editors August 17, 1998; accepted for publication (in revised form) June 10, 1999; published electronically June 22, 2000.
http://www.siam.org/journals/sima/32-1/34359.html
†Department of Mathematics, Northwestern University, Evanston, IL 60208 (clark@math.northwestern.edu).

one-dimensional map just as in the previous papers. The problem of the unfolding of the bifurcation is thus reduced to a question of understanding the unfolding of a certain type of one-dimensional map. In all the cases of the homoclinic bifurcation of the three-dimensional flow satisfying a set of assumptions, the resulting one-dimensional map can be shown to have a transitive invariant set for correctly chosen parameter values.

The standard symmetric untwisted situation leads to a symmetric one-dimensional problem which is monotonically increasing on both sides. In this paper, we consider one-dimensional maps which are not symmetric; in one case, the map is increasing on one side and decreasing on the other side. We present the results of the thesis of Byers [2], which show how to carry through the result of Williams [18] to show that the one-dimensional map is transitive in these nonsymmetric cases when the absolute value of the derivative is greater than square root of 2. We also refer to the recent result of Morales and Pujals [9], a previous work of Li and Yorke [8], and the thesis of Choi [3], which show that if the absolute value of the derivative is greater than one, then the map has a transitive invariant set which is not always the whole original interval. One example of these transitive invariant sets has a stable set which forms a dense open subset of a neighborhood, but the invariant set does not have a trapping region. We are interesting in verifying that the corresponding flow on $\mathbb{R}^3$ does have a trapping region so we give some conditions in section 2 which imply its existence.

A *trapping region* for a map f is a nonempty open set U such that the closure of the image of U is contained in the interior of U, $\mathrm{cl}(f(U)) \subset \mathrm{int}(U)$. A set Λ is called an *attracting set* provided there is a trapping region U such that $\Lambda = \bigcap_{k\geq 0} f^k(U)$. A set Λ is called an *attractor* provided it is an attracting set and $f|\Lambda$ is chain transitive. These definitions follow those given in [11].

For an attracting set Λ, there is a neighborhood U such that for any point $x \in U$ the ω-limit set of x is contained in Λ (i.e., the distance from $f^i(x)$ to Λ goes to zero as i goes to infinity); however, this condition is not equivalent to the existence of a trapping region. The simplest example is the map

$$f(x) = x + \frac{1}{2}x^2(1-x)^2 \qquad \text{for } x \bmod 1.$$

The ω-limit set of any point x_0 is 0, but there is no trapping region for $x = 0$. The problem is that the orbit of a small positive x_0 needs to go far away (near 0.5) before it returns near 1, which is equal to 0 mod 1.

There are other definitions of attractors, including Milnor's, which would call 0 an attractor for the map above. He only requires that there be a set B of positive measure such that the ω-limit sets of points in B are contained in Λ; i.e., B is contained in the stable set of Λ.

In this paper we consider another concept between our definition of an attractor and Milnor's: we call a set Λ a *weak attractor* provided (i) there is a neighborhood U of Λ and a dense open subset U' of U such that for all $x_0 \in U'$ the ω-limit set of x_0 is contained in Λ, and (ii) $f|\Lambda$ is chain transitive.

A weak attractor can have a 1-cycle in the terminology of Palis; i.e., there can be points $x_0 \in U \setminus \Lambda$ which are on both the stable and unstable set of Λ. If the map is f is one to one, then it has a 1-cycle provided there is a point $x_0 \notin \Lambda$ for which $\omega(x_0) \subset \Lambda$ and $\alpha(x_0) \subset \Lambda$; if f is not invertible, then it has a 1-cycle provided there is a point $x_0 \notin \Lambda$ for which $\omega(x_0) \subset \Lambda$ and there is some choice of preimages $\{x_i\}_{i\leq 0}$ with $f(x_{i-1}) = x_i$ for $i \leq 0$ and the distance from x_i to Λ goes to zero as i goes to $-\infty$. In the example given above, $x = 0$ is a weak attractor with a 1-cycle: for any

point $x_0 \in (0,1)$, $\alpha(x_0) = 0$ and $\omega(x_0) = 1 = 0 \bmod 1$. In section 2, we given another type of example of a map with a weak attractor but not an attractor in our sense of the term.

In this paper, as in [12] and [13], we consider a homoclinic bifurcation from the situation where there is a resonance between the eigenvalues together with transversality conditions. There are two other results by Rychlik [16] and Dumortier, Kokubu, and Oka [4] and Oka [10] which give different homoclinic bifurcations to Lorenz attractors than the one we analyze. These other authors assume there is no resonance of the eigenvalues, but each also assumes that there is a type of nontransversality along the homoclinic orbit (which is different in the two papers), while we assume there is transversality.

In section 2, we present the results on the one-dimensional maps. The main theorem about the homoclinic bifurcation of flows is given in section 3 together with the assumptions that are needed for this result. Section 4 contains the proof of the homoclinic bifurcation theorem. Section 5 contains some further comments about the unfolding of the bifurcation. Finally, section 6 proves that the assumptions for the bifurcation can be satisfied for specific polynomial differential equations in $\mathbb{R}^3$.

2. One-dimensional results. We are interested in conditions which imply that a one-dimensional map with a single discontinuity is topologically transitive.

We consider a map $f : J \to \mathbb{R}$ where $J \subset \mathbb{R}$ is an open interval and which we assume satisfies the following conditions:

(a) The map f has a discontinuity at a single point $c \in J$.

(b) The map f is continuously differentiable on $J \setminus \{c\}$, with

$$\lambda = \inf_{x \in J \setminus \{c\}} |f'(x)| > 1.$$

(c) The right and left limits of f exist at c: let

$$a^+ = \lim_{x \to c+} f(x) \quad \text{and} \quad a^- = \lim_{x \to c-} f(x).$$

Often we act as if f is not defined at c, but we could always take $f(c) = a^+$, $f(c) = a^-$, or $f(c) = c$.

We state the last two assumptions separately for the cases when f has the same monotonicity for x less than c and x greater than c. First, we consider the case when f is either monotonically increasing on both sides of c or monotonically decreasing on both sides.

(d1) Let $a = \max\{a^-, a^+\}$ and $b = \min\{a^-, a^+\}$. We assume that $b < c < a$, so that c is in the interior of the interval $[b, a]$.

(e1) Finally, we assume that $b < f(a)$, $f(b) < a$, so that the interval $[b, a]$ is invariant, $f([b, a]) \subset [b, a]$.

Next, we consider the case when f is monotonically increasing on one side of c and monotonically decreasing on the other side.

(d2) Let $a = \max\{a^-, a^+\}$ and $b = f(a)$. We assume that $b = f(a) < c < a$, so that c is in the interior of the interval $[b, a]$. (If $a = \min\{a^-, a^+\} < c$ and $b = f(a) > c$, then a reversal of orientation changes this case into the one considered here.)

(e2) Finally, we assume that $b < f(b) < a$, so that the interval $[b, a]$ is invariant.

It is not very hard to check that if f satisfies assumptions (a)–(e), then there is a small $\epsilon > 0$ such that the slightly larger interval $[b - \epsilon, a + \epsilon]$ is a trapping region. See Figure 2.1 for graphs of maps satisfying (a)–(e).

According to a theorem of Williams [18], if a map f satisfies conditions (a)–(e) and has the same monotonicities on both the subintervals $[b, c)$ and $(c, a]$, and $\lambda > \sqrt{2}$, then f is topologically transitive on $[a, b]$. Theorem 2.6 below gives a generalization of this result to other cases when f is increasing on one of the subintervals and is decreasing on the other.

There are other results that extend the results to the case of a map that satisfies conditions (a)–(e) for any $\lambda > 1$. Li and Yorke [8] proved that such maps have an ergodic measure whose support can be a subset of the original interval. More recently, Morales and Pujals [9] proved a different generalization of the result of Williams: they proved that if the map f satisfies conditions (a)–(e) for any $\lambda > 1$, then there is a closed subset $L_f \subset [b, a]$ which contains c in its interior such that f is topologically transitive on L_f and a dense open subset of points of $[b, a]$ have forward orbits which eventually are contained in L_f (the stable manifold of L_f is dense and open in $[b, a]$). In fact, L_f contains an interval I with c in its interior and L_f is the forward orbit of I. In general, the set L_f is the support of the measure found earlier by Li and Yorke.

In [3], Choi made more explicit the properties of L_f. In particular, (i) L_f is the finite union of closed intervals; (ii) the maximal invariant set in $\mathrm{cl}(J \setminus L_f)$ is a hyperbolic repeller R_f; (iii) L_f is always a weak attractor as defined in the introduction, but [3] gives an example where there is no trapping region for L_f so L_f is not an attractor in our strong sense of the term. The repeller R_f can be a set of periodic orbits and their preimages. (There are cases when R_f contains wandering points which have an α-limit set in one periodic orbit in R_f and an ω-limit set in another periodic orbit in R_f.) It is also possible for R_f to be a subshift of finite type as an example below shows. Choi has also shown that there are examples for which the set L_f does not have a trapping region (so L_f is not an attractor); such examples have a repelling periodic point on the boundary of L_f, i.e., a periodic orbit in $R_f \cap L_f$. For this example, the set L_f has a 1-cycle of the type discussed in the introduction. We give a different example below for which L_f does not have a trapping region, but without a 1-cycle. Choi also showed that the map can always be perturbed to a new map g without periodic points on the boundary of L_g, so L_g has a trapping region and so is an attractor for the new map g. We give a different example where L_h is not an attractor below.

We summarize these results in the following theorem.

THEOREM 2.1. *Assume that $f : J \to \mathbb{R}$ satisfies the assumptions* (a)–(e) *above with $\lambda > 1$.*

(a) *(Morales and Pujals) There is a $\delta_f > 0$ such that f is topologically transitive on*

$$L_f \equiv \mathrm{cl}\{\mathcal{O}^+((c - \delta_f, c + \delta_f), f)\},$$

and

$$W^s(L_f, f) \equiv \{x \in J : f^i(x) \in L_f \text{ for some } i \geq 0\}$$

is dense and open in J.

(b) *(Choi)* (i) *The set L_f is the finite union of closed intervals $\bigcup_{i=1}^n [x_i, y_i]$ and the endpoints*

$$\{x_i, y_i\}_{i=1}^n \subset \mathcal{O}^+(a^+, f) \cup \mathcal{O}^+(a^-, f).$$

(ii) *The maximal invariant set in $\mathrm{cl}(J \setminus L_f)$ is a closed hyperbolic repelling set R_f. (Some of the points in R_f can be wandering.)* (iii) *The set L_f has a trapping region*

for f if and only if

$$R_f \cap L_f = \emptyset, \quad \textit{i.e.},$$
$$\mathrm{Per}(f) \cap \partial(L_f) = \emptyset.$$

(iv) *Assume* $\mathrm{Per}(f) \cap \partial(L_f) \neq \emptyset$*, so f does not have a trapping region. There is a point $z_0 \in \mathrm{Per}(f) \cap \partial(L_f)$ with $f^j(z_0) = z_0$ and $f^k(a^\sigma) = z_0$, where σ is either $+$ or $-$. If g is near enough to f and $g^k(a_g^\sigma)$ does not have period j for g, $g^{j+k}(a_g^\sigma) \neq g^k(a_g^\sigma)$, then L_g will have a trapping region for g.*

We give some examples to clarify the theorem. When $L_f \neq J$, the maximal invariant set in the gaps R_f is often a collection of repelling periodic orbits and orbits whose α-limit set is one of these orbits and whose ω-limit set is another orbit. The first example gives an example where R_f is a single repelling fixed point; the second example shows that the set R_f can be a subshift of finite type.

Example 2.2. A simple example of a function f for which L_f is not the whole interval $[a, b]$ is given by

$$f(x) = \begin{cases} \dfrac{4}{3}\,x + 10 & \text{for } -6 \le x \le 0, \\ -1.3x + 10 & \text{for } 0 \le x \le 11. \end{cases}$$

Note that f is not differentiable at 0 so $c = 0$, $f(0) = 10$, $f(10) = -3$, $f(-3) = 6$, $f(6) = 2.2$, and $f(2.2) = 7.14 > 6$. Therefore $[a, b] = [-3, 10]$, and the transitive set $L_f = [-3, 2.2] \cup [6, 10]$. The set R_f is the single fixed point $10/2.3 \approx 4.35$.

Example 2.3. An example of a function g for which R_g is a subshift of finite type is given by

$$g(x) = \begin{cases} \dfrac{4}{3}\,x + 18 & \text{for } -21 \le x \le 0, \\ -\dfrac{7}{6}\,x + 18 & \text{for } 0 \le x \le 6, \\ -5\,(x - 6) + 11 & \text{for } 6 \le x \le 10, \\ -\dfrac{9}{8}\,(x - 10) - 9 & \text{for } 10 \le x \le 20. \end{cases}$$

The orbit of the nondifferentiable point 0 is not eventually periodic: $g(0) = 18$, $g(18) = -18$, $g(-18) = -6$, $g(-6) = 10$, $g(10) = -9$, $g(-9) = 6$, and $g(6) = 11 > 10$. The transitive set is $L_g = [-18, -9] \cup [-6, 6] \cup [10, 18]$. The repeller R_g is determined by the images of the gaps and is a subshift of finite type: $g([-9, -6]) = [6, 10]$, and $g([6, 10]) = [-9, 11] \supset [-9, -6] \cup [6, 10]$. Since none of the endpoints of L_g are periodic, $R_g \cap L_g = \emptyset$ and L_g has a trapping region.

Example 2.4. If we change the function g above so that $h(6) = 10$ but keep h piecewise linear with images of -18, -9, -6, 0, 10, and 18 unchanged, then $L_h = [-18, -9] \cup [-6, 6] \cup [10, 18]$ is the same as the last example but has a period three orbit on its boundary: $h^3(10) = h^2(-9) = h(6) = 10$. The repeller R_h is still a subshift of finite type, but $R_h \cap L_h = \{-9, 6, 10\} \neq \emptyset$ is a periodic orbit. The set L_h does not have a trapping region for h since it is accumulated upon by points in R_h outside of L_h. The stable set of L_h will include $[-19, 20] \setminus R_h$, which is dense and open in $[-19, 20]$ but is not a neighborhood of L_h. Therefore, L_h is a weak attractor but not an attractor (in terms of our terminology).

By approximating h by k with $k(6) = 10 + \epsilon > 10$, we can assure that L_k has a trapping region. Notice that we still have that $k^4(0) = 10$ but this is no longer a

periodic point and $L_k \cap \mathrm{Per}(k) = \emptyset$. Theorem 2.1(iv) states if this type of perturbation is possible for any map without a trapping region, then it can be approximated by a map with a trapping region. In terms of the application to homoclinic bifurcations of differential equations, this means that the differential equation needs to be able to be perturbed in such a way that the unstable manifold of the fixed point is not in the stable manifold of the given periodic orbit which gives the boundary orbit.

In the rest of this section, we give conditions from the thesis of Byers [2] which imply that L_f is the whole interval $[b, a]$. As stated above, Williams [18] proved this in the case when f has the same monotonicities. In the general results about a "Lorenz attractor" from a homoclinic bifurcation in Theorem 3.1, we only verify the weaker conditions of Theorem 2.1. In the "nearly symmetric" case where $C^+_{\eta_0} \approx C^-_{\eta_0}$, the results of Theorem 2.6 apply to show that there is a "Lorenz attractor" that has a one-dimensional return map with a single transitive interval.

Rather than prove directly that the map f is topologically transitive, we verify another condition called weakly locally eventually onto; Williams called a map $f : [b, a] \to [b, a]$ *locally eventually onto* (LEO) provided that for any nonempty open subinterval K there is an $n > 0$ such that $f^n(K) = [b, a]$. A map $f : [b, a] \to [b, a]$ is said to be *weakly locally eventually onto* (WLEO) provided that for any nonempty open interval $K \subset [b, a]$ there are an $n > 0$ and a finite set of points A such that $\bigcup_{i=0}^{n} f^i(K) = [b, a] \setminus A$; i.e., the forward orbit of K misses at most a finite set of points. It is easier to verify that a map is WLEO than LEO, and it still implies that the map is topologically transitive on $[b, a]$ by the Birkhoff transitivity theorem.

In proving that these maps f are WLEO, there are several cases depending on whether f is increasing or decreasing on the two subintervals $[b, c)$ and $(c, a]$. Graphs of these cases are given in Figure 2.1.

Case (i) (the original Lorenz map). The map f is increasing on both subintervals $[b, c)$ and $(c, a]$, $a = a^- > c$, $b = a^+ < c$, $b \le f(b)$, and $f(a) \le a$.

Case (ii) (the twisted Lorenz map). The map f is decreasing on both subintervals $[b, c)$ and $(c, a]$, $a = a^+ > c$, $b = a^- < c$, $f(b) \le a$, and $b \le f(a)$.

Case (iii) (variation of case (ii): the left endpoint is not a^- but is the image of a^+). The function f is decreasing on both subintervals $[b, c)$ and $(c, a]$, $a = a^+ > c$, $b = f(a) < c$, $b < a^-$, and $f(b) < a$.

Case (iv). The function f is increasing on $[b, c)$ and decreasing on $(c, a]$, $a = a^+ \ge a^- > c$, $b = f(a) < c$, and $b \le f(b)$.

Case (v). The function f is increasing for $[b, c)$ and decreasing on $(c, a]$, $a = a^- > a^+ > c$, $b = f(a) < c$, and $b \le f(b)$.

There are other cases with $a^+, a^- < c$ which are equivalent to Cases (iv) and (v) by a change of orientation which we do not list.

The proof of Williams [18] shows that in Cases (i)–(ii), if $\lambda > \sqrt{2}$ then f is LEO and transitive on all of $[b, a]$. As was shown in [2], for Cases (iv) and (v) this is not true: f is not always transitive on all of $[b, a]$ even when $\lambda > \sqrt{2}$. We state this in the following theorem.

THEOREM 2.5 (Byers). *Assume that $f : J \to \mathbb{R}$ satisfies the assumptions* (a)–(e) *above.*

In Case (iv) *above, there is a fixed point $p \in (c, a)$. Assume that $f(x) < p$ for all $x \in [b, c)$. Then f is not transitive on $[b, a]$.*

In Case (v) *above, there is an orbit of period two, $\{q^-, q^+\}$ with $q^- \in (b, c)$ and $q^+ \in (c, a)$. Assume that $c < a^+ < q^+$ and $q^- = f(q^+) < f(b) = f^2(a^-)$. Then f is not topologically transitive on $[b, a]$.*

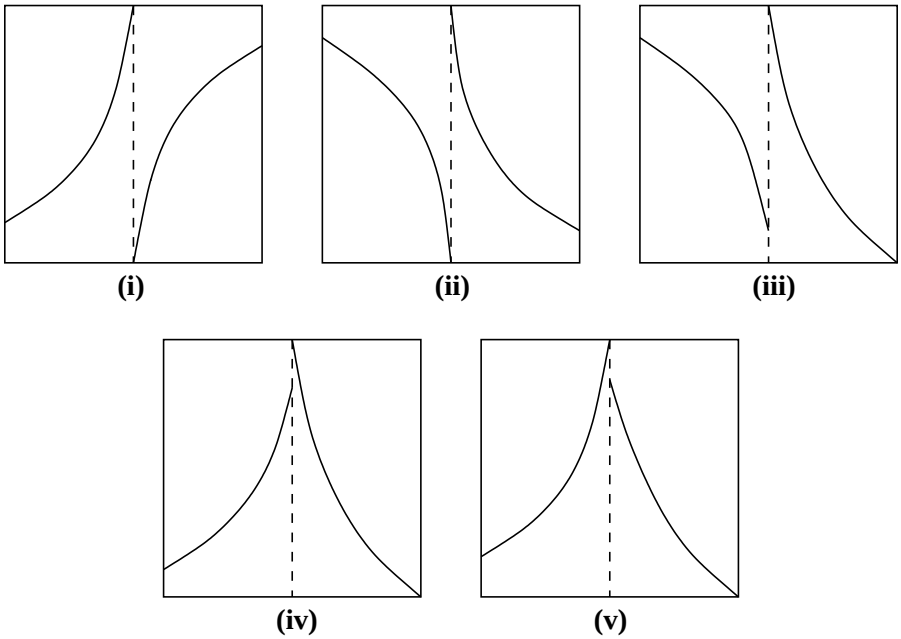


FIG. 2.1. *Graphs of Cases* (i)–(v).

Proof (idea of the proof).

Case (iv): Since $f'(x) > 1$ and $f[c, a] = [f(a), a]$, it follows that $f(a) < c$ and $f[c, a] \supset [c, a]$. Because the interval covers itself, there is a fixed point $p \in (c, a)$. (Notice that the fixed point cannot be either of the endpoints.)

This fixed point p must be repelling because $f'(x) > 1$ everywhere. There is an interval K about p which covers itself but is not in the image of any other points in $[b, a] \setminus K$. Therefore K is not contained in the transitive attractor L and f is not topologically transitive on all of $[b, a]$. See [2] for details.

An example of such a function given in [2] is

$$f(x) = \begin{cases} 1.6x + 0.35 & \text{for } -0.5 \le x < 0, \\ -1.5x + 1 & \text{for } 0 \le x \le 1. \end{cases}$$

Case (v): Let $\ell(K)$ be the length of an interval K. If $f(b) \ge c$, then $f[b, c] = [f(b), a] \subset [c, a]$ and $\ell(f[b, c]) \ge \lambda\ell([b, c])$. Then $f^2[b, c] = [b, f^2(b)]$ and $\ell(f^2[b, c]) \ge \lambda\ell(f[b, c]) \ge \lambda^2\ell([b, c])$, so this interval covers itself, $f^2[b, c] \supset [b, c]$. Thus there is a point of period two with $q^- \in [b, c]$, $q^+ = f(q^-) \in f[b, c] \subset [c, a]$. This proves the existence of a point of period two under the assumption that $f(b) \ge c$.

Otherwise, $f(b) < c$. We also have that $a^- > a^+ > c$. Then $f[b, c] \supset [c, a]$ and $f^2[b, c] \supset f[c, a] = [b, a^+] \supset [b, c]$. Again, there is a point of period two as desired.

With the assumptions of the theorem for f in Case (v), $q^+ > a^+$ and $q^- < f(b)$. Therefore there is a neighborhood K of $\{q^-, q^+\}$ made up of two intervals, which covers itself but is not in the image of any other points in $[b, a] \setminus K$. Therefore K is not contained in the transitive attractor L and f is not topologically transitive on all of $[b, a]$. See [2] for details.

An example of such a function given in [2] is

$$f(x) = \begin{cases} 1.415\,x + 1 & \text{for } -0.815 \le x \le 0, \\ -1.415\,x + 0.6 & \text{for } 0 \le x \le 1. \end{cases} \qquad \square$$

Notice that the examples given of the above theorem satisfy $\lambda > \sqrt{2}$ and are still not WLEO or topologically transitive. Therefore it is necessary to add further assumptions in order to insure that the map f is topologically transitive.

We now combine the various results in [2] into a single theorem.

THEOREM 2.6 (Byers). *Assume f satisfies assumptions* (a)–(e) *above and $\lambda > \sqrt{2}$. With the following added assumptions in each of the cases, $f|[a,b]$ is WLEO and so topologically transitive.*

In Cases (i) *or* (ii) *no added assumption is needed.*

In Case (iii), *further assume that $f(a^-) \ge p$, where $p \in (c,a)$ is the fixed point.*

In Case (iv), *where $a^+ \ge a^-$, further assume that $a^- \ge p$, where $p \in (c,a)$ is the fixed point.*

In Case (v), *where $a^- > a^+$, further assume that $a^+ \ge q^+$, where $q^+ \in (c,a]$ is the point of period two.*

Remark 2.7. Byers proved in [2] that in Case (iv) it is sufficient to assume that $(1+\sqrt{2})a^- \ge a^+$: this condition implies that $a^- \ge p$. Similarly, in Case (v) it is sufficient to assume that $(3-\sqrt{2})a^+ \ge a^-$; this condition implies that $a^+ \ge q^+$.

The proofs for all of the cases use the same basic construction due to Williams. Given an open interval $K \subset [b,a]$, we define inductively a sequence of intervals $K_i \subset [b,a]$ for $i \ge 0$. Define K_0 to be the longer component of $K \setminus \{c\}$. (Note that if $c \notin K$, then $K_0 = K$.) If K_j is defined for $0 \le j < i$, then let K_i be the longer component of $f(K_{i-1}) \setminus \{c\}$. Since $f(K_{i-1})$ is an open interval at each stage, it follows that all the K_i are open intervals.

Let $\ell(K)$ be the length of an open interval K.

LEMMA 2.8. *If $\lambda > \sqrt{2}$, then there exists an $n > 0$ such that $c \in f(K_{n-1})$ and $c \in f(K_n)$, so $c \in \partial(K_n) \cap f(K_n)$.*

Proof. If $c \notin f(K_i)$, then $\ell(K_{i+1}) \ge \lambda \ell(K_i)$. On the other hand, if $c \in f(K_i)$, then $c \in \partial(K_{i+1})$ and $\ell(K_{i+1}) \ge \frac{\lambda}{2}\ell(K_i)$. So if $c \notin f(K_{i-1}) \cap f(K_i)$ we get that

$$\ell(K_{i+1}) \ge \frac{\lambda^2}{2}\,\ell(K_{i-1}).$$

Since $\frac{\lambda^2}{2} > 1$, this cannot go on indefinitely, and there must be an $n > 0$ such that $c \in f(K_{n-1}) \cap f(K_n)$. $\square$

In the proofs below, we take n as given in the above lemma for which $c \in \partial(K_n) \cap f(K_n)$.

Proof (Case (i)). This is the case considered by Williams in [18]. We do not assume that $f(b) < c$ or $f(a) > c$. However, by modifying the argument in [18] or [11] in ways similar to the cases below, it still follows that f is WLEO. $\square$

Proof (Case (ii)). Because f expands lengths by a factor of $\lambda > 1$, it follows that $f(b) > c$ and $f(a) < c$. Therefore the proof is exactly as given before. $\square$

Proof (Case (iii)). If $K_n \subset (c,a]$, then $c \in \partial(K_n) \cap f(K_n)$ implies that

$$f(K_n) \supset [c,a) \quad \text{and} \quad f^2(K_n) \supset (b,a).$$

On the other hand, if $K_n \subset [b,c)$, then

$$
\begin{aligned}
&f(K_n) \supset (a^-, c],\\
&f^2(K_n) \supset (a^-, p] \supset [c, p],\\
&f^3(K_n) \supset [p, a^+) = [p, a), \quad \text{and}\\
&f^4(K_n) \supset (b, p].
\end{aligned}
$$

Therefore $f^3(K_n) \cup f^4(K_n) \supset (b, p] \cup [p, a) = (b, a)$. This completes the proof of this case. □

Proof (Case (iv)). This case is very similar to Case (iii). We leave the details to the reader. Also see [2]. □

Proof (Case (v)). If $K_n \subset [b, c)$, then

$$f(K_n) \supset [c, a^-) = [c, a), \quad \textit{and} \quad f^2(K_n) \supset (b, a^+) \supset (b, c].$$

Therefore $f(K_n) \cup f^2(K_n) \supset (b, a)$.

On the other hand, suppose $K_n \subset (c, a]$. Then

$$
\begin{aligned}
&f(K_n) \supset [c, a^+) \supset (c, q^+),\\
&f^2(K_n) \supset (f(a^+), a^+) \supset (q^-, c),\\
&f^3(K_n) \supset (q^+, a), \quad \textit{and}\\
&f^4(K_n) \supset (b, q^-).
\end{aligned}
$$

Therefore

$$f(K_n) \cup f^2(K_n) \cup f^3(K_n) \cup f^4(K_n) \supset (b, a) \setminus \{q^-, q^+\}.$$

This completes the proof of this case and the theorem. □

3. Statement of results for a homoclinic bifurcation. In this section we give the assumptions on flows in three dimensions which insure that a homoclinic bifurcation to a Lorenz attractor can take place. The first six assumptions, (A1)–(A6), on the parameterized differential equations concern the properties at the bifurcation value, η_0. The last assumption, (A7), is on the unfolding of the parameter η which ensures that there are parameter values that possess an attractor. The parameter space needs to be big enough to verify the assumptions of the one-dimensional map given in the last section.

(A1) We consider a C^2 vector field X_η on $\mathbb{R}^3$ which depends on the parameter η and which has a fixed point $\mathbf{Q}_\eta$ for all parameter values near η_0. We assume that the eigenvalues of $DX_\eta(\mathbf{Q}_\eta)$ are all real with $\lambda_{ss}(\eta) < \lambda_s(\eta) < 0 < \lambda_u(\eta)$ and with respective eigenvectors $\mathbf{v}^{ss}$, $\mathbf{v}^s$, and $\mathbf{v}^u$,

With this assumption, there are several invariant manifolds for the fixed point at the origin. We denote the one-dimensional unstable manifold tangent to $\mathbf{v}^u$ by $W^u(\mathbf{Q}_\eta, \eta)$ and the two-dimensional stable manifold tangent to the $\mathbf{v}^s$ and $\mathbf{v}^{ss}$ by $W^s(\mathbf{Q}_\eta, \eta)$. Next, there is a one-dimensional strong stable manifold tangent to $\mathbf{v}^{ss}$ which we denote by $W^{ss}(\mathbf{Q}_\eta, \eta)$. This latter manifold is made up of points which converge to $\mathbf{Q}_\eta$ at an asymptotic rate determined by the eigenvalue λ_{ss}. All of these manifolds are C^r if the vector field is C^r and are even real analytic if the vector field is real analytic. Finally, there is a two-dimensional manifold tangent to the two most expanding directions, $\mathbf{v}^u$ and $\mathbf{v}^s$, which we denote by $W^{eu}(\mathbf{Q}_\eta, \eta)$. This manifold

is local in the stable direction but can be extended along the unstable manifold by flowing forward in time. We call this the *extended unstable manifold* even though it is not expanding in all directions. (Some people call this the center unstable manifold.) This manifold is at least C^1 (and C^2 with assumption (A4) on the dominance of the contraction toward $W^{eu}(\mathbf{Q}_\eta, \eta)$ given by $e^{\lambda_{ss}}$ in comparison with the greatest contraction within $W^{eu}(\mathbf{Q}_\eta, \eta)$ given by e^{λ_s}.) With this notation we can make the second assumption about the existence of a homoclinic orbit. Without a symmetry assumption on the differential equations it is a codimension two condition to have a double homoclinic connection.

(A2) For the bifurcation value η_0, there is a *double homoclinic connection* with the unstable manifold of $\mathbf{Q}_{\eta_0}$ contained in the stable manifold but outside the strong stable manifold,

$$\Gamma \equiv W^u(\mathbf{Q}_{\eta_0}, \eta_0) \subset W^s(\mathbf{Q}_{\eta_0}, \eta_0) \setminus W^{ss}(\mathbf{Q}_{\eta_0}, \eta_0).$$

(The fact that Γ misses the strong stable manifold can be expressed as a transversality condition by stating that $W^u(\mathbf{Q}_{\eta_0}, \eta_0)$ is transverse to $W^{ss}(\mathbf{Q}_{\eta_0}, \eta_0)$.) In fact, we assume that the two branches $\Gamma^{\pm}$ of $\Gamma \setminus \{\mathbf{Q}_{\eta_0}\}$ are contained in the same component of $W^s(\mathbf{Q}_{\eta_0}, \eta_0) \setminus W^{ss}(\mathbf{Q}_{\eta_0}, \eta_0)$: $\Gamma = \{\mathbf{Q}_{\eta_0}\} \cup \Gamma^+ \cup \Gamma^-$.

(A3) For η_0, the two-dimensional extended unstable manifold $W^{eu}(\mathbf{Q}_{\eta_0}, \eta_0)$ is transverse to the two-dimensional stable manifold $W^s(\mathbf{Q}_{\eta_0}, \eta_0)$ along Γ.

The transversality condition in (A3) is generically satisfied and so does not add a codimension to the bifurcation. Let

$$P(\mathbf{q}) \equiv T_{\mathbf{q}} W^{eu}(\mathbf{Q}_{\eta_0}, \eta_0) \qquad \text{for } \mathbf{q} \in \Gamma.$$

Note that $P(\mathbf{Q}_{\eta_0})$ is spanned by $\mathbf{v}^u$ and $\mathbf{v}^s$. The transversality condition in (A3) together with the condition $W^u(\mathbf{Q}_{\eta_0}, \eta_0) \cap W^{ss}(\mathbf{Q}_{\eta_0}, \eta_0) = \emptyset$ in assumption (A2) implies that $P(\mathbf{q})$ converges to $P(\mathbf{Q}_{\eta_0})$ as $\mathbf{q}$ converges to $\mathbf{Q}_{\eta_0}$ along Γ by the inclination lemma (lambda lemma). Therefore $\{P(\mathbf{q}) : \mathbf{q} \in \Gamma\}$ is a continuous bundle over Γ. Considering one half of the homoclinic connection $\Gamma^+ \cup \mathbf{Q}_{\eta_0}$, let $\nu^+ = 1$ if the bundle $\{P(\mathbf{q}) : \mathbf{q} \in \Gamma^+ \cup \mathbf{Q}_{\eta_0}\}$ is orientable (not twisted) and $\nu^+ = -1$ if this bundle is nonorientable (twisted). In the same way considering the other half of the homoclinic connection $\Gamma^- \cup \mathbf{Q}_{\eta_0}$, let $\nu^- = \pm 1$ whenever the bundle $\{P(\mathbf{q}) : \mathbf{q} \in \Gamma^- \cup \mathbf{Q}_{\eta_0}\}$ is orientable or nonorientable, respectively. If the bundle is orientable, then the resulting one-dimensional map (which is discussed in the next section) is increasing on the corresponding subinterval; if the bundle is nonorientable then the resulting one-dimensional map is decreasing on the corresponding subinterval.

(A4) We assume that for η_0 the strong stable eigenvalue dominates the other two eigenvalues in the sense that

$$\lambda_{ss}(\eta_0) + [\lambda_u(\eta_0) - \lambda_s(\eta_0)] < 0 \quad \textit{and} \quad \lambda_{ss}(\eta_0) < 2\,\lambda_s(\eta_0).$$

This is an open condition and so does not add a codimension to the bifurcation. The second inequality in (A4) is what ensures that the manifold $W^{eu}(\mathbf{Q}_{\eta_0}, \eta_0)$ is C^2. See Theorem 5.1 in [6]. It is also redundant with the resonance assumption (A6) (although sometimes we want to assume (A4) but not necessarily assume (A6)). These conditions are used to prove that the one-dimensional Poincaré map is differentiable.

The next assumption on the equations is a restriction on the total change in area within the $P(q)$ directions ("within the attractor directions") when a solution travels the whole length of one of the loops Γ^+ or Γ^-.

(A5) Let $\mathbf{q}^{\pm}(t)$ be a parameterization of the solution along $\Gamma^{\pm}$. An initial vector $\mathbf{v}_0$ can be translated along the trajectory by the linearized equations (called the first variation equations in differential equations)

$$\dot{\mathbf{v}}(t) = DX_{\mathbf{q}^{\pm}(t)}\,\mathbf{v}(t)$$

to give a vector solution $\mathbf{v}(t)$ with $\mathbf{v}(0) = \mathbf{v}_0$. (If $\phi^t(\mathbf{x}_0)$ is the flow for the initial condition $\mathbf{x}_0$, then $\mathbf{v}(t) = D(\phi^t)_{\mathbf{x}_0}\mathbf{v}_0$.) One such solution is $X_\eta(\mathbf{q}^{\pm}(t))$. Let $\mathbf{v}^{\pm}(t)$ be such a solution of the linearized equations along $\mathbf{q}^{\pm}(t)$ so that $P(\mathbf{q}^{\pm}(t))$ is spanned by $\mathbf{v}^{\pm}(t)$ and $X_\eta(\mathbf{q}^{\pm}(t))$ for each t. Let $A^{\pm}(t)$ be the area of the parallelogram in $P(\mathbf{q}^{\pm}(t))$ spanned by $\mathbf{v}^{\pm}(t)$ and $X_\eta(\mathbf{q}^{\pm}(t))$. Define $C^{\pm}_{\eta_0}$ by

$$C^{\pm}_{\eta_0} = \lim_{t\to\infty} \frac{A^{\pm}(t)}{A^{\pm}(-t)}.$$

The quantity $C^{\pm}_{\eta_0}$ is the change in area within the planes $P(\mathbf{q}^{\pm})$ along the whole length of $\Gamma^{\pm}$. We assume that the limit exists and $0 < C^{\pm}_{\eta_0} < 1$. If $C^{+}_{\eta_0} \approx C^{-}_{\eta_0}$, then the interval $[b_\eta, a_\eta]$ in Lemma 4.2 is nearly symmetric about c_η and we only need $0 < C^{\pm}_{\eta_0} < 2$.

Assumption (A5) is an open condition.

Lemma 4.1 in the next section shows that $C^{\pm}_{\eta_0}$ has meaning in terms of a one-dimensional Poincaré map, f_{η_0}, as the coefficient of the lowest order nonconstant term. Therefore, $f'_{\eta_0}(c_{\eta_0}\pm) = \nu^{\pm}\,C^{\pm}_{\eta_0}$ in the sense of the one-sided limit of f' from above and below c_{η_0}. The fact that $C^{\pm}_{\eta_0} < 2$ means that f_{η_0} stretches lengths by a factor less than 2 and there is a hope that for η near η_0, f_η will map the appropriate interval $[b_\eta, a_\eta]$ inside itself (since there is one discontinuity). We restrict to $C^{\pm}_{\eta_0} < 1$ because in the proof this allows us to verify the necessary inequalities for $E_\eta = |\lambda_s(\eta)|/\lambda_u(\eta) < 1$. The fact that $C^{\pm}_{\eta} > 0$ makes it is possible for the derivative of the one-dimensional map to have absolute value greater than 1 in the desired interval for $\eta \neq \eta_0$. Lemma 4.2 gives conditions on unfolding parameters a^{+}_{η}, a^{-}_{η}, and $e_\eta = 1 - E_\eta = 1 - |\lambda_s(\eta)|/\lambda_u(\eta)$, which ensures that this interval is invariant and that the absolute value of the derivative is always bigger than 1.

In order to understand why the limit defining $C^{\pm}_{\eta_0}$ exists and to motivate the next assumption, we remember that in two dimensions the divergence gives the infinitesimal change of area. Let $\mathrm{div}_2(\mathbf{q}^{\pm}(t)) = A'(t)$ be the infinitesimal change of area within the two-dimensional planes $P(\mathbf{q}^{\pm}(t))$ as the solution $\mathbf{q}^{\pm}(t)$ moves along Γ, i.e., the "two-dimensional divergence in $P(\mathbf{q})$" along Γ. In terms of this quantity,

$$C^{\pm}_{\eta_0} = \exp\left(\int_{-\infty}^{\infty} \mathrm{div}_2(\mathbf{q}^{\pm}(t))\,dt\right).$$

The plane $P(\mathbf{q}^{\pm}(t))$ converges exponentially to the plane spanned by the eigenvectors $\mathbf{v}^s$ and $\mathbf{v}^u$ for the eigenvalues $\lambda_s(\eta_0)$ and $\lambda_u(\eta_0)$, so the quantity $\mathrm{div}_2(\mathbf{q}^{\pm}(t))$ converges exponentially to $\lambda_u(\eta_0) + \lambda_s(\eta_0)$. If $\lambda_u(\eta_0) + \lambda_s(\eta_0) \neq 0$, then $\mathrm{div}_2(\mathbf{q}^{\pm}(t)) \neq 0$ for $|t|$ large, the integral in expressing $C^{\pm}_{\eta_0}$ in terms of $\mathrm{div}_2(\mathbf{q}^{\pm}(t))$ would be $\pm\infty$, $C^{\pm}_{\eta_0}$ would be ∞ or 0, and the total change of area along $\Gamma^{\pm}$ would be ∞ or 0. On the other hand, if $\lambda_u(\eta_0) + \lambda_s(\eta_0) = 0$, then $\mathrm{div}_2(\mathbf{q}^{\pm}(t))$ converges exponentially to 0, the integral converges to a finite limit, and $C^{\pm}_{\eta_0}$ is a positive, nonzero quantity. Therefore, the final resonance assumption for η_0 makes assumption (A5) possible. This resonance condition is a codimension one condition; in total, the conditions of η_0

are codimension three. (Two codimensions are from the double homoclinic connection, and the resonance condition gives the third and final codimension.)

(A6) There is a one-to-one resonance between the unstable and weak stable eigenvalues for η_0:

$$\lambda_u(\eta_0) + \lambda_s(\eta_0) = 0.$$

Letting $E_\eta = |\lambda_s(\eta)|/\lambda_u(\eta)$ and $e_\eta = 1 - E_\eta$, this condition can be expressed by saying that $E_{\eta_0} = 1$ or $e_{\eta_0} = 0$.

The final assumption relates to the unfolding of the bifurcation.

(A7) We need to assume that the parameter space is big enough so that a_η^+, a_η^-, and $E_\eta = |\lambda_s(\eta)|/\lambda_u(\eta)$ can be varied independently for η near η_0. (If the equations are symmetric, as was the case in [12] and [13], then we need only assume that a_η^+ and E_η can be varied independently for η near η_0.)

It is now possible to state the main theorem.

THEOREM 3.1. *Assume that the vector field in $\mathbb{R}^3$, depending on a parameter η, is C^2 and satisfies assumptions* (A1)–(A7). *Let $\mathcal{N}$ be a small neighborhood of η_0 in parameter space. Then, there exists a subset $\mathcal{N}' \subset \mathcal{N}$ with nonempty interior such that $\eta_0 \in \mathrm{cl}(\mathcal{N}')$ and such that for $\eta \in \mathcal{N}'$ the flow for η has a topologically transitive weak attractor which contains the fixed point $\mathbf{Q}_\eta$. In fact, the weak attractor is determined by a one-dimensional Poincaré map f_η which is WLEO on a finite union of closed intervals L_η containing a single point of discontinuity in its interior. The values of $\nu^\pm$ determine whether the attractor is orientable or not on the two branches. If the vector field is C^3, then the resulting one-dimensional Poincaré map f_η for $\eta \in \mathcal{N}'$ has an ergodic invariant measure with support equal to the whole invariant set L_η and equivalent to Lebesgue on L_η.*

The proof of the theorem is contained in the next section.

Remark 3.2. The fact that the flow satisfies assumptions (A1)–(A4) means that standard stable manifold theory applies to show that the problem can be reduced to a one-dimensional Poincaré map. Thus, with the given assumptions, the proof of the theorem reduces to analyzing the unfolding of the one-dimensional map and showing that we can get the situation discussed in Theorem 2.1. The three unfolding parameters of the one-dimensional map are e_η and the two constant terms $a_\eta^\pm$ which are defined in Lemma 4.1. The proof indicates more fully what part of the parameter space yields an attractor. This is discussed more fully in section 5.

Remark 3.3. Although we call these Lorenz attractors for the differential equations, if the equations are very nonsymmetric (C_η^+ and C_η^- have very different values) then the one-dimensional Poincaré map will be transitive on a set made up of a finite number of intervals and not just one. In other words, the results of Morales–Pujals and Choi apply rather than Byers' extension of the result of Williams. Therefore all we verify is that the invariant set is a weak attractor. We believe that for a dense and open set of values $\eta \in \mathcal{N}'$, the invariant set is an attractor and not just a weak attractor. To prove this would require showing that by changing the parameters e_η and $a_\eta^\pm$, it is possible to realize the type of perturbations of the one-dimensional map f_η indicated in Theorem 2.1(b)(iv).

Remark 3.4. If the equations are nearly symmetric in the sense that

$$\sqrt{2} - 1 < \frac{C_{\eta_0}^+}{C_{\eta_0}^-} < \frac{1}{\sqrt{2} - 1},$$

then it is possible to insure that the derivative is greater than $\sqrt{2}$ in absolute value. For these parameters the equations have an attractor and the one-dimensional map is topologically transitive on a single interval I_η. Theorem 2.6 is only used to show that this remark is true and not in the proof of Theorem 3.1 as stated.

Remark 3.5. The existence of an ergodic invariant measure follows as in [13], using the result of Keller [7].

4. Proof of the theorem from the assumptions. We begin the proof by discussing the construction of the Poincaré map from the homoclinic connection and its form as given in [13].

Let Σ be a transversal to both $\Gamma^{\pm}$ out a short distance along the local stable manifold of $\mathbf{Q}_{\eta_0}$. The section can be taken with one component since we are assuming both branches of $\Gamma^{\pm}$ are in the same component of $W^s(\mathbf{Q}_{\eta_0}) \setminus W^{ss}(\mathbf{Q}_{\eta_0})$. There is a neighborhood $V \subset \Sigma$ of $\Gamma \cap \Sigma$ such that points in $V \setminus W^s(\mathbf{Q}_{\eta_0})$ return to Σ for $\eta = \eta_0$, defining a Poincaré map

$$F_{\eta_0} : V \setminus W^s(\mathbf{Q}_{\eta_0}, \eta_0) \subset \Sigma \to \Sigma.$$

Since the flow varies continuously with the parameter, the Poincaré map is defined for η near enough to η_0,

$$F_\eta : V \setminus W^s(\mathbf{Q}_\eta, \eta) \subset \Sigma \to \Sigma.$$

In [13], it was shown that assumption (A4) implies that the flow has an invariant continuous bundle of strong stable directions over Γ, $\{E^{ss}(\mathbf{q}) : \mathbf{q} \in \Gamma\}$, with $E^{ss}(\mathbf{Q}_{\eta_0}, \eta_0) = \langle \mathbf{v}^{ss} \rangle$. These conditions are open, so this bundle exists not only over Γ for η_0 but also over a neighborhood of Γ for nearby η. Then the stable manifold theory implies that there is a $C^{1+\mu}$ (C^1 plus μ-Hölder for some $\mu > 0$) invariant strong stable foliation in a neighborhood of Γ for η near η_0. If we take the union of these locally along an orbit and then intersect these with Σ, we get a one-dimensional foliation of Σ which is invariant by F_η. The projection along the leaves of the strong stable manifolds of orbits defines a map $\pi_\eta : \Sigma \to \Sigma^1$. By changing the orientation of Σ^1 if necessary, we can insure that we do not have $\nu^- = -1$ and $\nu^+ = 1$. (This last case can be changed into $\nu^- = 1$ and $\nu^+ = -1$.) The projection π_η can be used to define a one-dimensional map

$$f_\eta : V^1 \setminus \{c_\eta\} \subset \Sigma^1 \to \Sigma^1$$

by $f_\eta(\pi_\eta \mathbf{q}) = \pi_\eta F_\eta(\mathbf{q})$, where $V^1 = \pi_\eta(V)$ and $c_\eta = \pi_\eta(W^s(\mathbf{Q}_\eta, \eta) \cap V)$ is the point of discontinuity.

We need to analyze the one-dimensional map well enough to show that for correctly chosen parameter values it has a transitive invariant set containing the point of discontinuity. The next lemma, which was proved in [12] and [13], gives an expansion of the map which is used to prove the existence of such a set. First we label the constant terms of the expansion of f_η; let

$$a_\eta^{\pm} = \limsup_{\tau \to c_\eta \pm} f_\eta(\tau).$$

This quantity corresponds to the signed distance of $\Gamma_\eta^{\pm} \subset W^u(\mathbf{Q}_\eta, \eta)$ from $W^s(\mathbf{Q}_\eta, \eta)$ as measured in Σ^1.

LEMMA 4.1. *Assume assumptions* (A1)–(A4) *are satisfied. Let* E_η *and* $C_{\eta_0}^{\pm}$ *be defined as in assumptions* (A6) *and* (A5). *Let* $c_\eta = \pi(W^s(\mathbf{Q}_\eta, \eta) \cap V)$. *Let* $J \subset \Sigma^1$

be a fixed small interval about c_{η_0}. For η in a small neighborhood of η_0, the induced one-dimensional Poincaré map $f_\eta : J \setminus \{c_\eta\} \subset \Sigma^1 \to \Sigma^1$ has continuous derivative on $J \setminus \{c_\eta\}$, and f_η and f'_η have the following form:

$$f_\eta(\tau) = \begin{cases} a_\eta^+ + \nu^+ C_\eta^+ |\tau - c_\eta|^{E_\eta} + o(|\tau - c_\eta|^{E_\eta}) & \text{for } \tau > c_\eta, \\ a_\eta^- - \nu^- C_\eta^- |\tau - c_\eta|^{E_\eta} + o(|\tau - c_\eta|^{E_\eta}) & \text{for } \tau < c_\eta, \end{cases}$$

$$f'_\eta(\tau) = \begin{cases} \nu^+ E_\eta C_\eta^+ |\tau - c_\eta|^{E_\eta - 1} + o(|\tau - c_\eta|^{E_\eta - 1}) & \text{for } \tau > c_\eta, \\ \nu^- E_\eta C_\eta^- |\tau - c_\eta|^{E_\eta - 1} + o(|\tau - c_\eta|^{E_\eta - 1}) & \text{for } \tau < c_\eta. \end{cases}$$

The constants $C_\eta^\pm$ depend continuously on η.

See [13] for the lemma's proof. The proof uses either (i) linearization near the fixed point or (ii) the analysis of the flow past the fixed point using a normal form of the vector field. Also see [14].

Let $a_\eta = \max\{a_\eta^-, a_\eta^+\}$ and $b_\eta = \min\{a_\eta^-, a_\eta^+, f(a_\eta)\}$. We are interested in parameter values η for which $b_\eta < c_\eta < a_\eta$. In order to have an expanding attractor for these parameter values, we also need f_η to preserve the interval $[b_\eta, a_\eta]$ and the absolute value of the derivative to be greater than 1 for points in the interval.

The three unfolding parameters that we use are $a_\eta^\pm$ and e_η. The parameters $a_\eta^\pm$ measure the extent to which the homoclinic connections are broken (and to which sides). The quantity $e_\eta = 1 - E_\eta = 1 - |\lambda_s(\eta)|/\lambda_u(\eta)$ measures the extent to which the two eigenvalues are no longer in resonance.

For the three allowable cases of $\nu^\pm$, if we take parameter values η for which $\nu^+(a_\eta^+ - c_\eta) < 0$ and $\nu^-(a_\eta^- - c_\eta) > 0$, then $a_\eta > c_\eta$.

Lemma 4.2 proves that η_0 can be approximated by parameter values η for which $f_\eta([b_\eta, c_\eta)) \subset [b_\eta, a_\eta]$, $f_\eta((c_\eta, a_\eta]) \subset [b_\eta, a_\eta]$, $|f'_\eta(a_\eta)| > 1$, and $|f'_\eta(b_\eta)| > 1$. Since $|f'_\eta(a_\eta)| > 1$ and $|f'_\eta(b_\eta)| > 1$, the form of f'_η given in Lemma 4.1 implies that there is a $\lambda > 1$ such that $|f'_\eta(\tau)| \geq \lambda$ for all $\tau \in [b_\eta, a_\eta]$. By the result of Morales and Pujals [9] summarized in Theorem 2.1(a), this implies that there is a transitive invariant set L_{f_η} containing c_η which is a weak attractor. Because the one-dimensional map can be varied by changing the flow, if there is a periodic point on the boundary of L_{f_η} then it seems likely that it can be perturbed away. If this is indeed the case, then by the results of [3] summarized in Theorem 2.1(b), either L_η is an attractor for f_η or η can be perturbed to η' for which $L_{\eta'}$ is an attractor for $f_{\eta'}$. Because of the relationship between the flow and the one-dimensional Poincaré map, this shows that the flow for η has a transitive weak attractor as claimed in the theorem. Most likely it can be approximated by a parameter η' which has a transitive attractor as discussed in Remark 3.3. The claim about the ergodic measure for the one-dimensional map follows just as in [13] using the result of Keller [7]. Thus we only need to prove the following lemma.

LEMMA 4.2. *Let*

$$\begin{aligned} \mathcal{N}' = \{\eta \in \mathcal{N} : e_\eta > 0,\ \nu^+(a_\eta^+ - c_\eta) < 0,\ \nu^-(a_\eta^- - c_\eta) > 0, \\ f([b_\eta, c_\eta)) \subset [b_\eta, a_\eta],\ f((c_\eta, a_\eta]) \subset [b_\eta, a_\eta], \\ |f'_\eta(a_\eta)| > 1,\ |f'_\eta(b_\eta)| > 1\ \}. \end{aligned}$$

Then $\mathcal{N}' \neq \emptyset$ and $\eta_0 \in \mathrm{cl}(\mathcal{N}')$. The conditions on e_η and $a_\eta^\pm$ which ensure that $\eta \in \mathcal{N}'$ are given by inequality (4.1) *below when $\nu^+ = \nu^- = 1$ and by inequality* (4.3) *when $\nu^+ = \nu^- = -1$.*

Remark 4.3. When $\nu^+ = \nu^- = \pm 1$, the interval found for $\eta \in \mathcal{N}'$ extends from a_η^- to a_η^+ and satisfies

$$\frac{\log|a_\eta^+ - c_\eta|}{\log|a_\eta^- - c_\eta|} \approx \frac{\log(C_\eta^+)}{\log(C_\eta^-)}.$$

Note that for $C_\eta^- \neq C_\eta^+$, the interval is not symmetric about c_η.

When $\nu^- = -\nu^+ = 1$, the parameters found for $\eta \in \mathcal{N}'$ satisfy $a_\eta^+ \approx a_\eta^-$. One end of the interval is $a_\eta = \max\{a_\eta^+, a_\eta^-\}$, and the other end is $b_\eta = f_\eta(a_\eta)$. In the proof below we show that

$$\max\left\{0, \frac{C_\eta^+}{C_\eta^-} - 1\right\} < \frac{c_\eta - b_\eta}{a_\eta - c_\eta} < \frac{C_\eta^+}{C_\eta^-}.$$

Again, the interval is not symmetric about c_η when $C_\eta^- \neq C_\eta^+$.

Proof. First consider the case when $\nu^+ = \nu^- = 1$. These maps fall into Case (i). To get a transitive attractor, we take parameter values such that $a_\eta = a_\eta^- > c_\eta$ and $b_\eta = a_\eta^+ < c_\eta$. If $C_\eta^+ = C_\eta^-$ (as in the symmetric case), we can take $a_\eta^+ - c_\eta \approx -(a_\eta^- - c_\eta)$; this is the situation considered in Lemma 2 of [13]. We need to allow C_η^+ to have a value very different from C_η^- even though both are in the interval $(0, 2)$. We want the derivative to be greater than 1:

$$\begin{aligned} &1 < |f_\eta'(a_\eta)| \approx E_\eta C_\eta^+ |a_\eta^- - c_\eta|^{E_\eta - 1} \quad \textit{and} \\ &1 < |f_\eta'(b_\eta)| \approx E_\eta C_\eta^- |a_\eta^+ - c_\eta|^{E_\eta - 1}. \end{aligned}$$

Also, we need the interval $[b_\eta, a_\eta]$ to be invariant, so

$$\begin{aligned} &|a_\eta^- - c_\eta| + |c_\eta - a_\eta^+| \geq |f_\eta(a_\eta^-) - a_\eta^+| \approx C_\eta^+ |a_\eta^- - c_\eta|^{E_\eta} \quad \textit{and} \\ &|a_\eta^- - c_\eta| + |c_\eta - a_\eta^+| \geq |f_\eta(a_\eta^+) - a_\eta^-| \approx C_\eta^- |a_\eta^+ - c_\eta|^{E_\eta}. \end{aligned}$$

Thus the conditions on a_η^+, a_η^-, and e_η are approximately the following:

$$\begin{aligned} -\log(E_\eta C_\eta^-) < &e_\eta \log(|a_\eta^+ - c_\eta|^{-1}) < \log\left(1 + \left|\frac{a_\eta^- - c_\eta}{a_\eta^+ - c_\eta}\right|\right) - \log(C_\eta^-), \\ -\log(E_\eta C_\eta^+) < &e_\eta \log(|a_\eta^- - c_\eta|^{-1}) < \log\left(1 + \left|\frac{a_\eta^+ - c_\eta}{a_\eta^- - c_\eta}\right|\right) - \log(C_\eta^+). \end{aligned} \tag{4.1}$$

Since E_η goes to 1 as η goes to η_0, these conditions can be solved at the same time, with

$$\frac{\log(|a_\eta^+ - c_\eta|^{-1})}{\log(|a_\eta^- - c_\eta|^{-1})} \approx \frac{\log(C_\eta^-)}{\log(C_\eta^+)}. \tag{4.2}$$

When both $C_{\eta_0}^+, C_{\eta_0}^- < 1$, the resulting value of e_η can be made positive. If $C_{\eta_0}^+ \approx C_{\eta_0}^-$, then the interval is nearly symmetric, $\log\left(1 + \left|\frac{a_\eta^- - c_\eta}{a_\eta^+ - c_\eta}\right|\right) \approx \log(2)$, and we need only $C_{\eta_0}^\pm < 2$.

Next consider the case when $\nu^+ = \nu^- = -1$. These maps could fall into Cases (ii) or (iii), but we just use Case (ii) to show that parameter values exist. Again the

symmetric case was considered in [13]. In the general (possibly nonsymmetric) case, we choose parameter values so that $a_\eta = a_\eta^+ > c_\eta$ and $b_\eta = a_\eta^- < c_\eta$. We want

$$\begin{aligned} 1 < |f'_\eta(a_\eta)| &\approx E_\eta C_\eta^+ |a_\eta^+ - c_\eta|^{E_\eta - 1} \quad \textit{and} \\ 1 < |f'_\eta(b_\eta)| &\approx E_\eta C_\eta^- |a_\eta^- - c_\eta|^{E_\eta - 1}. \end{aligned}$$

Also, we need the interval $[b_\eta, a_\eta]$ to be invariant, so

$$\begin{aligned} |a_\eta^- - c_\eta| + |c_\eta - a_\eta^+| \geq |f_\eta(a_\eta^+) - a_\eta^+| &\approx C_\eta^+ |a_\eta^+ - c_\eta|^{E_\eta} \quad \textit{and} \\ |a_\eta^- - c_\eta| + |c_\eta - a_\eta^+| \geq |f_\eta(a_\eta^-) - a_\eta^-| &\approx C_\eta^- |a_\eta^- - c_\eta|^{E_\eta}. \end{aligned}$$

Thus the conditions on a_η^+, a_η^-, and e_η are approximately the following:

$$\begin{aligned} -\log(E_\eta C_\eta^+) < & e_\eta \log(|a_\eta^+ - c_\eta|^{-1}) < \log\left(1 + \left|\frac{a_\eta^- - c_\eta}{a_\eta^+ - c_\eta}\right|\right) - \log(C_\eta^+), \\ -\log(E_\eta C_\eta^-) < & e_\eta \log(|a_\eta^- - c_\eta|^{-1}) < \log\left(1 + \left|\frac{a_\eta^+ - c_\eta}{a_\eta^- - c_\eta}\right|\right) - \log(C_\eta^-). \end{aligned} \tag{4.3}$$

Again when both $C_\eta^+, C_\eta^- < 1$, these conditions can be solved at the same time with $e_\eta > 0$ and

$$\frac{\log(|a_\eta^+ - c_\eta|^{-1})}{\log(|a_\eta^- - c_\eta|^{-1})} \approx \frac{\log(C_\eta^+)}{\log(C_\eta^-)}. \tag{4.4}$$

We could enlarge the set of allowable parameter values to include those which give one-dimensional maps that fall into Case (iii) as long as $a_\eta^- - f_\eta(a_\eta^+)$ is small enough.

Finally, we consider the case when $\nu^+ = -1$ and $\nu^- = 1$. (The case with $\nu^+ = 1$ and $\nu^- = -1$ can be reduced to this case by reversing orientation of Σ^1.) These maps fall into Cases (iv) or (v). We first take parameter values η so that $a_\eta = a_\eta^- = a_\eta^+ > c_\eta$. After we obtain the result in this case, the interval remains invariant with absolute value of the derivative greater than 1 for $|a_\eta^+ - a_\eta^-|$ small and $a_\eta = \max\{a_\eta^-, a_\eta^+\}$.

Set $b_\eta = f_\eta(a_\eta)$. Thus $f_\eta[c_\eta, a_\eta] = [b_\eta, a_\eta]$ maps inside the relevant interval. We need to check that $|f'_\eta(a_\eta)| > 1$, $|f'_\eta(b_\eta)| > 1$, and $f_\eta(b_\eta) > b_\eta$. The last condition ensures that $f_\eta[b_\eta, a_\eta] \subset [b_\eta, a_\eta]$.

The derivative at a_η must satisfy

$$1 < |f'_\eta(a_\eta)| \approx E_\eta C_\eta^+ |a_\eta - c_\eta|^{-e_\eta}$$

or

$$-\log(C_\eta^+ E_\eta) < e_\eta \log(|a_\eta - c_\eta|^{-1}).$$

Next, we need $f_\eta(b_\eta) > b_\eta$. Since

$$\begin{aligned} f_\eta(b_\eta) - b_\eta &\approx f_\eta(a_\eta - C_\eta^+ (a_\eta - c_\eta)^{E_\eta}) - a_\eta + C_\eta^+ (a_\eta - c_\eta)^{E_\eta} \\ &\approx a_\eta - C_\eta^- [-a_\eta + C_\eta^+ (a_\eta - c_\eta)^{E_\eta} + c_\eta]^{E_\eta} - a_\eta + C_\eta^+ (a_\eta - c_\eta)^{E_\eta} \\ &= (a_\eta - c_\eta)^{E_\eta} \{C_\eta^+ - C_\eta^- [C_\eta^+ (a_\eta - c_\eta)^{-e_\eta} - 1]^{E_\eta}\}, \end{aligned}$$

$f_\eta(b_\eta) - b_\eta > 0$ provided

$$\frac{C_\eta^+}{C_\eta^-} > [C_\eta^+(a_\eta - c_\eta)^{-e_\eta} - 1]^{E_\eta}.$$

Since E_η converges to 1, this is approximately the inequality

$$\frac{C_\eta^+}{C_\eta^-} > C_\eta^+(a_\eta - c_\eta)^{-e_\eta} - 1 \quad or \quad \frac{1}{C_\eta^-} + \frac{1}{C_\eta^+} > (a_\eta - c_\eta)^{-e_\eta}.$$

This is the second inequality we need to satisfy. Thus these two conditions are satisfied provided (approximately)

$$\log\left(\frac{1}{C_\eta^+ E_\eta}\right) < e_\eta \log(|a_\eta - c_\eta|^{-1}) < \log\left(\frac{1}{C_\eta^-} + \frac{1}{C_\eta^+}\right). \tag{4.5}$$

These two inequalities can be satisfied at the same time.

To check that $|f'_\eta(b_\eta)| > 1$, we need to consider two subcases: (i) $1 > C_{\eta_0}^- > C_{\eta_0}^+ > 0$ and (ii) $1 > C_{\eta_0}^+ \geq C_{\eta_0}^- > 0$. First, define the comparison of the lengths of the two sides of the interval $[b_\eta, a_\eta]$ by

$$\gamma_\eta = \frac{c_\eta - f_\eta(a_\eta)}{a_\eta - c_\eta} = \frac{c_\eta - b_\eta}{a_\eta - c_\eta}.$$

We see below that $0 < \gamma_\eta < 1$ for subcase (i) and $\gamma_\eta > -1 + C_\eta^+/C_\eta^-$ for subcase (ii), which can often be greater than 1. Using Lemma 4.1 and the definition of γ_η,

$$\begin{aligned} a_\eta - b_\eta &\approx C_\eta^+(a_\eta - c_\eta)^{E_\eta}, \\ a_\eta - b_\eta &= a_\eta - c_\eta + c_\eta - b_\eta \\ &= (1 + \gamma_\eta)(a_\eta - c_\eta), \quad \text{so} \\ (1 + \gamma_\eta) &\approx C_\eta^+(a_\eta - c_\eta)^{-e_\eta}. \end{aligned}$$

For the first subcase (i) when $1 > C_{\eta_0}^- > C_{\eta_0}^+ > 0$, using (4.5),

$$\begin{aligned} 1 + \gamma_\eta &\approx C_\eta^+(a_\eta - c_\eta)^{-e_\eta}, \\ 1 < C_\eta^+(a_\eta - c_\eta)^{-e_\eta} &< 1 + \frac{C_\eta^+}{C_\eta^-}, \quad \text{so} \\ 0 < \gamma_\eta &< \frac{C_\eta^+}{C_\eta^-} < 1. \end{aligned}$$

Therefore, for parameters satisfying (4.5),

$$\begin{aligned} |f'_\eta(b_\eta)| &\approx E_\eta C_\eta^- |b_\eta - c_\eta|^{-e_\eta} \\ &\approx E_\eta C_\eta^- \gamma_\eta^{-e_\eta} |a_\eta - c_\eta|^{-e_\eta} \\ &> \frac{C_\eta^-}{C_\eta^+} \gamma_\eta^{-e_\eta} \\ &> \frac{C_\eta^-}{C_\eta^+} > 1, \end{aligned}$$

since $\gamma_\eta^{-1} > 1$. Thus for this subcase all three conditions are satisfied for parameters satisfying (4.5).

Now consider subcase (ii) when $1 > C^+_{\eta_0} \geq C^-_{\eta_0} > 0$. Again

$$|f'_\eta(b_\eta)| \approx E_\eta C^-_\eta \gamma_\eta^{-e_\eta} |a_\eta - c_\eta|^{-e_\eta}.$$

Inequality (4.5) together with the fact that $1 + \gamma_\eta \approx C^+_\eta (a_\eta - c_\eta)^{-e_\eta}$ imply that γ_η is bounded as η converges to η_0 in $\mathcal{N}'$. Therefore $E_\eta \gamma_\eta^{-e_\eta}$ converges to 1 as η converges to η_0 in $\mathcal{N}'$. Therefore the inequality $|f'_\eta(b_\eta)| > 1$ is (essentially) equivalent to

$$-\log(C^-_\eta) < e_\eta \log(|a_\eta - c_\eta|^{-1}).$$

Since $-\log(C^+_\eta) < -\log(C^-_\eta)$, this implies that all three conditions are satisfied in this subcase provided

$$\log\left(\frac{1}{C^-_\eta}\right) < e_\eta \log(|a_\eta - c_\eta|^{-1}) < \log\left(\frac{1}{C^-_\eta} + \frac{1}{C^+_\eta}\right). \tag{4.6}$$

Notice that for these parameters

$$\frac{C^+_\eta}{C^-_\eta} < C^+_\eta (a_\eta - c_\eta)^{-e_\eta} \approx 1 + \gamma_\eta < 1 + \frac{C^+_\eta}{C^-_\eta}, \quad \text{so}$$
$$0 \leq \frac{C^+_\eta}{C^-_\eta} - 1 < \gamma_\eta < \frac{C^+_\eta}{C^-_\eta},$$

which can be quite large if the system is very asymmetrical. This completes the proof of the lemma and theorem. □

5. Unfolding of the bifurcation. To make the discussion simpler, we assume that $c_\eta \equiv 0$. We assume that $0 < C^\pm_{\eta_0} < 1$ since this is the situation that leads to an attractor of Lorenz type. In fact, the situation we verify for specific equations in this paper and the papers [12], [13] has $0 < C^\pm_{\eta_0} << 1$. We discuss the cases $\nu^+ = \nu^- = \pm 1$ and $\nu^- = -\nu^+$ separately. In each case we take the relationship between a^+_η and a^-_η found in the proof of Lemma 4.2 so there are only two parameters, $e_\eta = 1 - E_\eta$ and either a^+_η or a^-_η

First consider $\nu^+ = \nu^- = 1$. By (4.2), $a^+_\eta \approx -|a^-_\eta|^\kappa$, where $\kappa = \log(C^-_\eta)/\log(C^+_\eta)$. So, we can use the two parameters $a_\eta = a^-_\eta$ and $e_\eta = 1 - E_\eta$.

The region of parameters labeled (ii) in Figure 5.1 is the region $\mathcal{N}'$ found in Theorem 3.1, which corresponds to systems with an attractor of Lorenz type. As a consequence of inequality (4.1) in the proof of Lemma 4.2, the boundary of $\mathcal{N}'$ is contained in $\partial\mathcal{N}$, γ_1, and γ_2, where the latter two are given approximately by

$$\gamma_1 : \quad e_\eta \log(|a^-_\eta|^{-1}) \approx \log\left(\frac{1}{C^+_\eta}\right), \quad a^-_\eta > 0,$$
$$\gamma_2 : \quad e_\eta \log(|a^-_\eta|^{-1}) \approx \log\left(\frac{1}{C^+_\eta}\right) + \log(1 + |a^-_\eta|^{\kappa-1}), \quad a^-_\eta > 0.$$

Notice that in the symmetric case $\kappa = 1$ and

$$\gamma_2 : \quad e_\eta \log(|a^-_\eta|^{-1}) \approx \log\left(\frac{2}{C^+_\eta}\right),$$

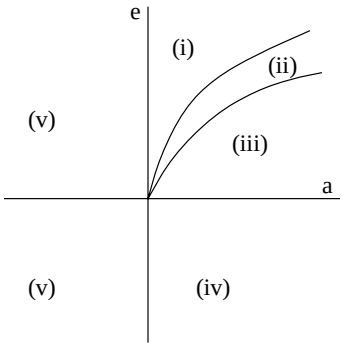


FIG. 5.1. *Bifurcation diagram.*

which is the form of the boundary given in [12] and [13].

Region (iii) in Figure 5.1: In this case the absolute value of the derivative is not greater than 1 at all points in $[a_\eta^+, a_\eta^-]$. However, for many of these parameters, the map is still eventually expanding and there is a transitive attractor. We do not attempt to analyze these cases more thoroughly.

Region (i) in Figure 5.1: In this region $a_\eta^- > 0$ and e_η is above γ_2. It follows that (a) $|f(a_\eta^-)| > a_\eta^-$, (b) the interval $[a_\eta^+, a_\eta^-]$ is not invariant, and (c) there is a horseshoe that is separated from the fixed point of the flow.

Region (iv) in Figure 5.1: In this case there is an attracting periodic orbit. See Remark 5.1 below.

Region (v) in Figure 5.1: Because $a_{\eta_0}^- < 0$, the discontinuity 0 is not in the image of the map, so there is no invariant set that bifurcates off near the homoclinic orbit.

The case for $\nu^- = \nu^+ = -1$ is very similar. By (4.4), $a_\eta^- \approx -|a_\eta^+|^\kappa$, where the exponent $\kappa = \log(C_\eta^-)/\log(C_\eta^+)$. So, we can use the two parameters $a_\eta = a_\eta^+$ and $e_\eta = 1 - E_\eta$. The interval is $[a_\eta^-, a_\eta^+]$.

Again the region of parameters labeled (ii) in Figure 5.1 is the region $\mathcal{N}'$ found in Theorem 3.1, which corresponds to systems with an attractor of Lorenz type. As a consequence of inequality (4.3) in the proof of Lemma 4.2, the boundary of $\mathcal{N}'$ is contained in $\partial\mathcal{N}$, γ_1, and γ_2, where the latter two are given approximately by

$$\gamma_1 : \quad e_\eta \log(|a_\eta^+|^{-1}) \approx \log\left(\frac{1}{C_\eta^+}\right), \quad a_\eta^+ > 0,$$

$$\gamma_2 : \quad e_\eta \log(|a_\eta^+|^{-1}) \approx \log\left(\frac{1}{C_\eta^+}\right) + \log(1 + |a_\eta^+|^{\kappa-1}), \quad a_\eta^+ > 0.$$

The other regions are similar to the previous case.

Finally we consider the case when $\nu^- = -\nu^+ = 1$. In this case we take $a_\eta = a_\eta^+ = a_\eta^-$. By inequalities (4.5) and (4.6), the two boundary components of region (ii) are now given approximately by

$$\gamma_1 : \quad e_\eta \log(|a_\eta^+|^{-1}) \approx \max\left\{\log\left(\frac{1}{C_\eta^+}\right), \log\left(\frac{1}{C_\eta^-}\right)\right\}, \quad a_\eta > 0,$$

$$\gamma_2 : \quad e_\eta \log(|a_\eta^+|^{-1}) \approx \log\left(\frac{1}{C_\eta^+} + \frac{1}{C_\eta^-}\right), \quad a_\eta^+ > 0, \quad a_\eta > 0.$$

The other regions are very similar to the situation of the previous cases.

Remark 5.1. Rovella [15] showed that there are flows with $E_\eta > 1$ and $\nu^\pm = 1$ that have transitive attractors. Such an attractor has a one-dimensional Poincaré map whose derivative is zero at the discontinuity. Rovella showed that the method of Benedicks and Carleson [1] could be used to show that there is a transitive attractor for a positive set of parameter values. These results should also apply when $\nu^\pm = -1$, but the case for $\nu^- = -\nu^+$ is very different and it is not clear that this argument applies.

These attractors do not occur in our unfolding for $\nu^- = \nu^+ = 1$ because (in the symmetric case) in order for $f_\eta(a_\eta^+) > c_\eta$ it is necessary for $C_\eta^+ > 1$ and we consider only $0 < C_\eta^+, C_\eta^- < 1$. On the other hand, if $C_{\eta_0}^+ = C_{\eta_0}^- > 1$ and $\nu^+ = \nu^- = \pm 1$, it seems likely that an attractor of the type found by Rovella occurs in the unfolding.

6. Specific differential equations satisfying the assumptions. In the previous papers [12] and [13] we showed that there were symmetric differential equations satisfying assumptions (A1)–(A7) with $\nu^+ = \nu^- = 1$ or $\nu^+ = \nu^- = -1$. In this section, we show that there is a polynomial differential equation satisfying (A1)–(A7) with $\nu^+ = -1$ and $\nu^- = 1$. The basic idea of the example is the same as before, but the equations need to be modified so that the twisting is different on the two sides. Because of the difference, it is no longer possible to make the equations have symmetry. Most of the verification of the assumptions is very straightforward. The two things that need to be checked more carefully are the transversality of assumption (A3) and the bound on the coefficients $C_{\eta_0}^\pm$ in assumption (A5).

The equations we consider are

$$\text{(NSE)} \qquad \begin{aligned} \dot{x} &= y, \\ \dot{y} &= x - 2\,x^3 - \alpha\,y + \beta\,x^2 y + \epsilon\,x^3 y + xyz, \\ \dot{z} &= -\gamma\,z + \delta\,x^2. \end{aligned}$$

The parameters are $\eta = (\alpha, \beta, \gamma, \delta, \epsilon)$. The changes from the equations considered in the previous papers [12] and [13] are that in the $\dot{y}$ equation we have added the term $\epsilon\,x^3 y$ and the term xyz replaces yz in [12] and xz in [13].

The fixed point $\mathbf{0}$ is always the origin. The linearization of the vector field is given by

$$DX_{(x,y,z)} = \begin{pmatrix} 0 & 1 & 0 \\ 1 - 6x^2 + 2\beta xy + 3\epsilon x^2 y + yz & -\alpha + \beta x^2 + \epsilon x^3 + xz & xy \\ 2\delta x & 0 & -\gamma \end{pmatrix}.$$

At the origin, the eigenvalues are $\lambda_{ss} = -\alpha/2 - (1 + \alpha^2/4)^{1/2}$, $\lambda_u = -\alpha/2 + (1 + \alpha^2/4)^{1/2}$, and $\lambda_s = -\gamma$, giving assumption (A1).

By picking the parameter $\gamma_0 = \lambda_u = -\alpha_0/2 + (1 + \alpha_0^2/4)^{1/2}$ at the bifurcation, we can insure that $\lambda_s(\eta_0) + \lambda_u(\eta_0) = 0$, giving assumption (A6).

To obtain (A4), we need the combination of all three eigenvalues less than zero, $\lambda_{ss}(\eta_0) - \lambda_s(\eta_0) + \lambda_u(\eta_0) < 0$:

$$\begin{aligned} 0 &> [-\alpha_0/2 - (1 + \alpha_0^2/4)^{1/2}] + 2[-\alpha_0/2 + (1 + \alpha_0^2/4)^{1/2}] \\ &> -3\alpha_0/2 + (1 + \alpha_0^2/4)^{1/2} \end{aligned}$$

or

$$\alpha_0 > 2^{-\frac{1}{2}}.$$

Thus to obtain a C^1 foliation, it is not possible to take a small perturbation of the integrable case where $\alpha = \beta = \delta = 0$. Given the resonance condition (A6), the second inequality in (A4) follows from the first.

To verify assumptions (A2), (A3), and (A5), we start with $\delta_1 = 0$. By adjusting the parameter values β_1 and ϵ_1 we can make a double homoclinic connection with $\delta_1 = 0$. For $\delta_1 = 0$ the (x, y)-plane is invariant and $W^u(\mathbf{0}, \eta_1) \subset W^{ss}(\mathbf{0}, \eta_1)$, so (A2) is not true. Just as in [13], we can perturb δ_1 to $\delta_0 > 0$ and adjust β_1 to β_0 and ϵ_1 to ϵ_0 to keep the double homoclinic connection. Because the $\delta_0 x^2$ term is positive in $\dot{z}$, the unstable manifold $W^u(\mathbf{0}, \eta_0)$ is pushed upward and $W^{ss}(\mathbf{0}, \eta_0)$ is pushed downward, giving assumption (A2), $W^u(\mathbf{0}, \eta_0) \cap W^{ss}(\mathbf{0}, \eta_0) = \emptyset$. Following the argument in [13], our Lemma 6.1 proves that after this perturbation with $\delta_0 > 0$ but small, the transversality condition of assumption (A3) is satisfied and that $0 < C_{\eta_0} < 2$ as required in assumption (A5). It also proves that $\nu^+ = -1$ and $\nu^- = -1$.

The unfolding assumption (A7) is satisfied because changing γ varies e_η while β and ϵ can adjust $a_\eta^\pm$ independently. Thus all that is left to prove is the following lemma.

LEMMA 6.1. *For $\delta_0 > 0$ but small, $W^{eu}(\mathbf{0}, \eta_0)$ is transverse to $W^s(\mathbf{0}, \eta_0)$, $0 < C_{\eta_0}^\pm << 1$, and $\nu^+ = -1$ and $\nu^- = -1$.*

Proof. A normal vector to $W^{eu}(\mathbf{0}, \eta)$, or $P(\mathbf{q}_\eta^\pm(t))$, is a covector and satisfies the adjoint differential equation

$$\dot{\mathbf{p}} = -\mathbf{p}\, DX(\mathbf{q}_\eta^\pm(t)).$$

(Note that in this equation, $\mathbf{p}$ is written as a row vector.) We denote the solution that is perpendicular to $P(\mathbf{q}_\eta^\pm(t))$ by $\mathbf{p}_\eta^\pm(t) = (p_1^\pm(t, \eta), p_2^\pm(t, \eta), p_3^\pm(t, \eta))$. Note that together, $(\mathbf{q}_\eta^\pm(t), \mathbf{p}_\eta^\pm(t))$ lies on the unstable manifold of $(\mathbf{0}, \mathbf{0})$ in the space of positions and covectors, $T^*\mathbb{R}^3$.

We start with $\delta_1 = 0$ and β_1 and ϵ_1 adjusted so that there are double homoclinic orbits. As t goes to infinity, we want to show that $p_3^\pm(t, \eta_1)$ goes to $-\infty$ and $\mathbf{p}_{\eta_1}^\pm(t)$ approaches the direction given by $-\mathbf{v}_s^*$ along the negative z-axis. The equations for $\dot{p}_1$ and $\dot{p}_2$ are independent of p_3 and so can be solved independently for a solution $(p_1^\pm(t, \eta_1), p_2^\pm(t, \eta_1))$ that is perpendicular to the homoclinic orbit in the (x, y)-plane, and so it limits on the eigendirection $\mathbf{v}_u^*$ for the eigenvalue $-\lambda_u$. Therefore $(p_1(t, \eta_1), p_2(t, \eta_1)) \to \mathbf{0}$ as t goes to infinity.

We parameterize the homoclinic connections $\mathbf{q}_\eta^\pm(t)$ so that $y^\pm(0, \eta) = 0$, so $x^\pm(t, \eta) y^\pm(t, \eta)$ is positive for $t < 0$ and negative for $t > 0$. Since

$$\begin{aligned}
\dot{p}_3^\pm(t, \eta) &= -x^\pm(t, \eta)\, y^\pm(t, \eta)\, p_2^\pm(t, \eta) + \gamma\, p_3^\pm(t, \eta), \\
p_3^\pm(t, \eta) &= e^{\gamma(t-t_0)} p_3^\pm(t_0, \eta) - \int_{t_0}^{t} x^\pm(s, \eta)\, y^\pm(s, \eta)\, p_2^\pm(s, \eta)\, e^{\gamma(t-s)}\, ds.
\end{aligned}$$

As t_0 goes to $-\infty$, since $(\mathbf{q}_\eta^\pm(t_0), \mathbf{p}_\eta^\pm(t_0))$ is in the unstable manifold in $T^*\mathbb{R}^3$, $\mathbf{p}_\eta^\pm(t_0)$ goes to zero exponentially at a rate given by λ_{ss}, $e^{-|\lambda_{ss}||t_0|}$. Since this is faster decay than the growth given by γ, we can let $t_0 \to -\infty$ and obtain

$$p_3^\pm(t, \eta) = -\int_{-\infty}^{t} x^\pm(s, \eta)\, y^\pm(s, \eta)\, p_2^\pm(s, \eta)\, e^{\gamma(t-s)}\, ds.$$

Because $X(\mathbf{q}_{\eta_1}^\pm(0))$ points in the direction of $(0, 1, 0)$, $p_2^\pm(0, \eta_1) = 0$ and we can take $p_2^\pm(t, \eta_1) > 0$ for $t < 0$ and $p_2^\pm(t, \eta_1) < 0$ for $t > 0$, so $x^\pm(t, \eta_1) y^\pm(t, \eta_1) p_2^\pm(t, \eta_1) \geq 0$

and

$$-p_3(t,\eta_1) > e^{\gamma t}\int_0^t x^{\pm}(s,\eta_1)\, y^{\pm}(s,\eta_1)\, p_2(s,\eta_1)\, e^{-\gamma s}\, ds.$$

As t goes to infinity, the integral is positive and $e^{\gamma t}$ goes to infinity, so $-p_3(t,\eta_1)$ goes to infinity. Since we showed above that $(p_1(t,\eta_1),p_2(t,\eta_1)) \to \mathbf{0}$, $\mathbf{p}^{\pm}(t,\eta_1)$ has a limiting direction along the negative z-axis, i.e., in the direction of the coeigenvector $-\mathbf{v}_s^*$ for the eigenvalue $-\lambda_s$. Therefore the limiting plane $P(\mathbf{q}^{\pm}_{\eta_1}(\infty))$ is spanned by the directions $\{\mathbf{v}^u, \mathbf{v}^{ss}\}$, and $W^{eu}(\mathbf{0},\eta_1)$ is transverse to $W^s(\mathbf{0},\eta_1)$. Since transversality is an open condition, it remains true for η_0 with $\delta_0 > 0$ small, giving assumption (A3).

We now want to show that $\nu^+ = -1$ and $\nu^- = 1$. The limiting direction of $\mathbf{p}^{\pm}_{\eta_0}(t)$ as t goes to $-\infty$ is $-\mathbf{v}^*_{ss}$, with a negative first coordinate. To understand the behavior as t goes to ∞, notice that for η_1, the limiting direction of $\mathbf{p}^{\pm}_{\eta_1}(t)$ is $-\mathbf{v}^*_s$, which is contained in the space spanned by $\{\mathbf{v}^*_{ss}, \mathbf{v}^*_s\}$. Since it is an open condition not to have a component in the $\mathbf{v}^*_u$ direction for the eigenvalue $-\lambda_u$, it will continue to be true for η_0. (This is the openness of the transverse intersection of assumption (A3).) For η_0 with $\delta_0 > 0$ but small, there is still a homoclinic connection, but now the homoclinic orbit $\mathbf{q}^{\pm}_{\eta_0}(t)$ approaches $\mathbf{0}$ along the weak stable direction. Since $\mathbf{p}^{\pm}_{\eta_0}(t)$ must remain orthogonal to X, the limiting direction of $\mathbf{p}^{\pm}_{\eta_0}(t)$ is contained in the space spanned by $\{\mathbf{v}^*_{ss}, \mathbf{v}^*_u\}$. Combining the two arguments, the limiting direction must be in the $\pm\mathbf{v}^*_{ss}$ direction. (This shows that the bundle of planes $\{P(\mathbf{q}) : \mathbf{q} \in \Gamma\}$ is continuous, and so is a second way of seeing that assumption (A3) is true.) Because the trajectory bends upward and then down asymptotic to the z-axis, the limiting direction of $\mathbf{p}^{+}_{\eta_0}(t)$ is $\mathbf{v}^*_{ss}$, while that of $\mathbf{p}^{-}_{\eta_0}(t)$ is $-\mathbf{v}^*_{ss}$. Therefore $\nu^+ = -1$ and $\nu^- = 1$.

As argued in [12], for η_1 with $\delta_1 = 0$, the integral of assumption (A5) is $-\infty$ and $C^{\pm}_{\eta_1} = 0$. By the perturbation argument given in [12], for $\delta_0 > 0$ small, the integral is still very negative but finite, so $0 < C^{\pm}_{\eta_0} << 1$. This proves assumption (A5). Notice that since we do not calculate the integrals, we have no way of knowing whether $C^{+}_{\eta_0}$ is nearly equal to $C^{-}_{\eta_0}$. □

REFERENCES

[1] M. Benedicks and L. Carleson, *On iterations of* $1 - ax^2$ *on* $(-1, 1)$, Ann. of Math. (2), 122 (1985), pp. 1–25.

[2] M. Byers, *Topologically Transitivity of a Class of Piecewise Monotone Expanding Maps of the Interval with a Single Discontinuity*, Thesis, Northwestern University, Evanston, IL, 1995.

[3] Y. Choi, *One Dimensional Lorenz-Like Attractors*, Thesis, Northwestern University, Evanston, IL, 1998.

[4] F. Dumortier, H. Kokubu, and H. Oka, *A degenerate singularity generating geometric Lorenz attractors*, Ergodic Theory Dynam. Systems, 15 (1995), pp. 833–856.

[5] R. Ghrist, P. Holmes, and M. Sullivan, *Knots and Links in Three-Dimensional Flows*, Lecture Notes in Math. 1654, Springer-Verlag, New York, 1997.

[6] M. W. Hirsch, C. C. Pugh, and M. Shub, *Invariant Manifolds*, Lecture Notes in Math. 583, Springer-Verlag, New York, 1977.

[7] G. Keller, *Generalized bounded variation and applications to piecewise monotonic transformations*, Z. Wahrsch. Verw. Gebiete, 69 (1985), pp. 461–478.

[8] T. Y. Li and J. Yorke, *Ergodic transformations from an interval into itself*, Trans. Amer. Math. Soc., 235 (1978), pp. 183–192.

[9] C. A. Morales and E. R. Pujals, *Singular strange attractors on the boundary of Morse-Smale Systems*, Ann. Sci. École Norm. Sup. (4), 30 (1997), pp. 693–717.

[10] H. Oka, *Bifurcation of Lorenz-like attractors from a degenerate vector field singularity*, in Towards the Harnessing of Chaos, M. Yamaguti, ed., The Seventh TOYOTA Conference Proceedings, Elsevier, New York, 1994, pp. 389–392.

[11] C. Robinson, *Dynamical Systems, Stability, Symbolic Dynamics, and Chaos*, CRC Press, Boca Raton, FL, 1995.
[12] C. Robinson, *Homoclinic bifurcation to a transitive attractor of Lorenz type*, Nonlinearity, 2 (1989), pp. 495–518.
[13] C. Robinson, *Homoclinic bifurcation to a transitive attractor of Lorenz type*, II, SIAM J. Math. Anal., 23 (1992), pp. 1255–1268.
[14] R. Roussarie, *On the number of limit cycles which appear by perturbation of separatrix loop of planar vector fields*, Bol. Soc. Brasil. Mat., 17 (1986), pp. 67–101.
[15] A. Rovella, *The dynamics of perturbations of the contracting Lorenz attractor*, Bol. Soc. Brasil. Mat. (N.S.), 24 (1993), pp. 233–259.
[16] M. Rychlik, *Lorenz attractors through Sil'nikov-type bifurcation, Part* I, Ergodic Theory Dynam. Systems, 10 (1990), pp. 793–822.
[17] N. Tufillaro, T. Abbott, and J. Reilly, *An Experimental Approach to Nonlinear Dynamics and Chaos*, Addison-Wesley, Reading, MA, 1992.
[18] R. F. Williams, *The structure of Lorenz attractors*, Inst. Hautes Études Sci. Publ. Math., 50 (1979), pp. 73–99.

SIAM J. MATH. ANAL.
Vol. 32, No. 1, pp. 142–171

ANALYSIS OF A FREE BOUNDARY PROBLEM ARISING IN BUBBLE DYNAMICS*

ZS. BIRO† AND J. J. L. VELAZQUEZ‡

Abstract. In this paper we study a model for the dynamics of a gas bubble immersed in a liquid. We describe the expansion of the bubble due to the decrease of the fluid pressure induced by a sound wave that traverses the fluid. The model considered here takes into account the dissipation of energy induced by the thermal conductivity. The expansion of the bubble can take place in a runaway manner in some cases. We obtain some global existence results for the model under consideration and we also describe a possible blow-up mechanism.

Key words. free boundary problems, bubble dynamics, parabolic differential equations, blow-up

AMS subject classifications. 35B40, 76N15

PII. S0036141099351681

1. Introduction.

In this paper we analyze a free boundary problem that arises in the study of the dynamics of gas bubbles in a liquid under the effect of an external sound field.

There exists a huge amount of literature concerning the dynamics of a gas bubble immersed in a liquid. It was predicted long time ago by Rayleigh [Ra] that, due to the interaction with an external sound wave, gas bubbles could experience in some circumstances a large expansion followed by an abrupt collapse. During the collapse of the bubble many interesting phenomena can occur, for instance, the breaking of symmetry of the bubble, sonochemistry, sonoluminiscence, and others [Le].

We are basically concerned with the analysis of the expansion stage of the process. In the models that are commonly used to describe bubble dynamics, it is usually assumed that the expansion part of the process is isothermal and that the collapse is approximately adiabatic [Le]. In any case, it is commonplace in several models to assume polytropic laws for the equation of state that relates the pressure and the density of the gas. Under these assumptions, together with the hypothesis of radial symmetry, it follows that the evolution of the bubble can be modeled by an ordinary differential equation (ODE) known as the Rayleigh–Plesset equation that will be described later. This ODE has been extensively studied using analytical and numerical methods (see [Le] and references therein).

We shall consider in this paper a model of bubble dynamics in which the effect of thermal conductivity of the gas is taken into account. This model arises in a natural way if the free boundary problem for the Navier–Stokes system that describes the evolution of the bubble in the radially symmetric case is written in nondimensional units. More precisely, we have made nondimensional the different magnitudes using the characteristic quantities that appear in some recent experimental studies of sonoluminiscence [BP1], [BP2], [BWLRP]. In the resulting model that we have derived

*Received by the editors February 1, 1999; accepted for publication (in revised form) November 23, 1999; published electronically June 22, 2000.
http://www.siam.org/journals/sima/32-1/35168.html

†Computer and Automation Institute, Hungarian Academy of Sciences, H-1518, Budapest, POB 63, Hungary. The research of this author was partially supported by Research Fund OTKA F-025464.

‡Departamento de Matemática Aplicada, Facultad de Matemáticas, Universidad Complutense, Madrid 28040, Spain (velazque@sunma4.mat.ucm.es). The research of this author was partially supported by DGICYT grant PB96-0614.

in this way (neglecting small nondimensional numbers) the pressure of the gas and the density are not related by a simple power law. Actually, the evolution of the gas during the expansion phase cannot be considered as either isothermal or isoentropic. It turns out, however, that both magnitudes (pressure and density) are related to the temperature by means of the usual equation of state for ideal gases, and on the other hand, the temperature satisfies a parabolic equation. The consequence of this analysis is that the whole problem for bubble evolution becomes a free boundary problem instead of an ODE problem.

The main goal of this paper is to analyze this free boundary problem during the expansion phase of the bubble. In the model that we consider a parameter Γ appears that measures the importance of inertial effects on the fluid compared with the changes of pressure induced by the external sound field. This parameter is very small during the expansion stage of the process. However, inertial effects play an essential role after the explosive growth of the radius of the bubble and are responsible for its collapse. For this reason, we have let the parameter Γ be different from zero in part of our analysis, since this quantity should become crucial after the runaway growth of the bubble.

In this article we show that the equilibrium state that corresponds to a constant radius is stable under small perturbations for the whole model with $\Gamma \neq 0$. To this end we use a suitable thermodynamic potential as a Lyapunov function. In the rest of the paper we assume $\Gamma = 0$ and we prove that all the solutions are global in time if the pressure of the liquid is always positive. On the other hand, using matched asymptotic expansions, we describe a blow-up mechanism that can take place for the model if the pressure of the liquid can become negative for some range of times. We have not attempted to describe in this paper the precise way in which inertial effects become relevant after the expansion of the bubble or the collapse of its radius.

2. Derivation of the model. In order to describe the dynamics of the bubble, we need to consider the motion of the liquid produced by the variation of size of the bubble. The dynamics of the liquid is described by the incompressible Navier–Stokes equation. We restrict our analysis to radially symmetric deformations of the bubble. The continuity equation for the liquid can then be written as

$$\frac{1}{r^2}\frac{\partial}{\partial r}(r^2 v_r) = 0, \qquad r > R(t), \tag{2.1}$$

where $v_r(r,t)$ is the radial velocity of the liquid and $R(t)$ is the radius of the bubble. Assuming that the fluid far away from the bubble is at rest, we obtain

$$v_r(r,t) = \frac{a(t)}{r^2}. \tag{2.2}$$

The momentum equation for the liquid is given under our symmetry hypothesis by

$$\frac{\partial v_r}{\partial t} = \frac{1}{\varrho_l}\Delta v_r - v_r\frac{\partial v_r}{\partial r} - \frac{1}{\varrho_l}\frac{\partial p_l}{\partial r}, \qquad r > R(t), \tag{2.3}$$

where ϱ_l, p_l are, respectively, the density and the pressure of the liquid. Plugging (2.2) into (2.3), it is readily seen that the viscous term disappears. We then have that

$$\dot{a}(t)\frac{1}{r^2} = \frac{2(a(t))^2}{r^5} - \frac{1}{\varrho_l}\frac{\partial p_l}{\partial r}, \qquad r > R(t). \tag{2.4}$$

Integrating (2.4) in the region $r > R(t)$ we obtain

$$\frac{\dot{a}(t)}{R(t)} = \frac{1}{2}\frac{(a(t))^2}{(R(t))^4} - \frac{1}{\varrho_l}p_\infty(t) + \frac{1}{\varrho_l}p_l(t), \tag{2.5}$$

where $p_l(t) = p_l(R(t), t)$ is the pressure of the liquid at the surface of the bubble and p_∞ is the liquid pressure far away from the bubble.

Taking into account the fact that the velocity of the liquid coincides with the speed of the bubble at its boundary, we deduce that

$$a(t) = (R(t))^2\dot{R}(t), \tag{2.6}$$

and using (2.6) in (2.5) we obtain the equation [Le]

$$R\ddot{R} + \frac{3}{2}(\dot{R})^2 = \frac{p_l(t) - p_\infty(t)}{\varrho_l}. \tag{2.7}$$

We have to complement (2.7) with the equations that describe the gas dynamics. These are the Navier–Stokes equations for compressible fluids that in the radially symmetric case read as

$$\frac{\partial \varrho}{\partial t} + \frac{1}{r^2}\frac{\partial}{\partial r}(\varrho r^2 v) = 0, \tag{2.8}$$

$$\varrho\left(\frac{\partial v}{\partial t} + v\frac{\partial v}{\partial r}\right) = \nu\Delta v - \frac{\partial p}{\partial r}, \tag{2.9}$$

$$\varrho T\frac{Ds}{Dt} = \nabla\cdot(\kappa\nabla T), \tag{2.10}$$

where ϱ is the density of the liquid, $p(r,t)$ is the pressure, $v(r,t)$ is the radial velocity, ν is the viscosity coefficient, κ is the thermal conductivity, T is the temperature, and s is the entropy per unit of mass. The definition of the convective derivative $\frac{D}{Dt}$ is the usual one:

$$\frac{D}{Dt} = \frac{\partial}{\partial t} + v\frac{\partial}{\partial r}.$$

We need to complement these equations with suitable equations of state. Taking into account the order of magnitude for the pressure, densities, and temperature that is reached inside the bubble during the expansion stage, it is enough to use the classical equations of state for ideal gases, namely,

$$p = R_g\varrho T, \tag{2.11}$$

$$s = c_v\log\left(\frac{p}{\varrho^\gamma}\right), \tag{2.12}$$

where $\gamma = \frac{c_p}{c_v}$ takes the well-known values $\frac{5}{3}$ for monoatomic gases and $\frac{7}{5}$ for diatomic gases.

Throughout this paper, we will assume that the temperature and density inside the bubble, as well as the remaining physical magnitudes, have the order of magnitude given by the experimental data in [BP1], [BP2], [BWLRP]. For the reader's convenience, we have included in Appendix A at the end of the paper a list of the sizes of the different quantities that are relevant for our problem. Let us denote as t_c the characteristic time associated to the frequency of oscillation, and let c be the order of magnitude of the sound speed inside the bubble. Let R_0 be the characteristic length that measures the size of the bubble. In our current situation (cf. Appendix A),

$$R_0 \ll ct_c.$$

Since the sound velocity is so large, disturbances in the fluid propagate through the bubble very quickly. In particular, writing the momentum equation (2.9) in nondimensional form we obtain to the lowest order

$$\frac{\partial p}{\partial r} = 0,$$

which implies

$$p = p(t). \tag{2.13}$$

On the other hand, the jump of the pressure at the boundary of the bubble is given by the surface tension. We are using here the fact that in our case the contributions due to viscosity terms are negligible compared with pressure forces. Then

$$p(t) = p(R(t), t) = p_l(t) + \frac{2\sigma}{R(t)}, \tag{2.14}$$

where from now on $p(t)$ denotes the pressure of the liquid. To conclude, notice that the nonslip condition on the side of the gas yields

$$v(R(t), t) = \dot{R}(t). \tag{2.15}$$

In several analyses of bubble dynamics, the gas pressure and its density are related by means of a power law. For instance, this approach can be found in [Le], where a polytropic law is assumed as an equation of state. The effect of the vapor pressure has also been taken into account there. Under these assumptions, the following relation between pressure and velocity holds:

$$p_l(t) = \left(p_0 + \frac{2\sigma}{R_0} - p_v\right)\left(\frac{R_0}{R}\right)^{3k} + p_v - \frac{2\sigma}{R(t)} \tag{2.16}$$

for some suitable k. Combining (2.7) and (2.16), the following differential equation results:

$$R\ddot{R} + \frac{3}{2}(\dot{R})^2 = \frac{1}{\varrho_l}\left\{\left(p_0 + \frac{2\sigma}{R_0} - p_v\right)\left(\frac{R_0}{R}\right)^{3k} + p_v - p_\infty(t) - \frac{2\sigma}{R(t)}\right\}. \tag{2.17}$$

Equation (2.17) is the well-known Rayleigh–Plesset equation that has been extensively studied by different authors (see [Le] for an extensive review). In this paper we will not assume a polytropic law as (2.16), but on the contrary we will relate p and ϱ

by means of (2.11) and introduce an additional equation for T. Since the thermal conductivity of the gas in nondimensional units is very high (see Appendix A), we can assume that the temperature of the liquid is a constant T_l and then coincides with the temperature of the gas at the surface of the bubble:

$$T(R(t),t) = T_l. \tag{2.18}$$

Our goal is to analyze the model (2.7), (2.8), (2.10)–(2.15). To this end we nondimensionalize the problem (see Appendix A) and plug (2.14) into (2.7). For the sake of convenience we will keep the same symbols to denote nondimensional quantities. The problem then becomes

$$R\ddot{R} + \frac{3}{2}(\dot{R})^2 = \frac{1}{\Gamma}\left(p(t) - p_\infty(t) - \frac{2\sigma}{R(t)}\right), \tag{2.19}$$

$$\frac{\partial \varrho}{\partial t} + \frac{1}{r^2}\frac{\partial}{\partial r}(\varrho r^2 v) = 0, \qquad B_{R(t)}(0) \times \mathbb{R}^+, \tag{2.20}$$

$$p = p(t), \qquad B_{R(t)}(0) \times \mathbb{R}^+, \tag{2.21}$$

$$\varrho T \frac{Ds}{Dt} = \nabla \cdot (\kappa \nabla T), \qquad B_{R(t)}(0) \times \mathbb{R}^+, \tag{2.22}$$

complemented with (2.11), (2.12), (2.14), and (2.18). The nondimensional number Γ in (2.19) is defined as

$$\Gamma = \frac{p_0 t_c^2}{R_0^2 \varrho_l},$$

where p_0 is the order of magnitude of the external pressure. In view of the data in Appendix A the parameter Γ turns out to be very small:

$$\Gamma \sim \frac{1}{400}.$$

The pressure field p_∞ that describes the variations of pressure due to the external sound field is given (in nondimensional units) by

$$p_\infty(t) = 1 + \alpha \cos(\omega t). \tag{2.23}$$

In the experiments of [BP1], [BP2], [BWLRP], α is approximately 1.2. Notice that this implies that during some time p_∞ becomes negative. In fact, these negative values of the pressure are responsible for the large expansion experienced by the bubble.

3. Some preliminary results. In this section we establish local well-posedness for the problem (2.11), (2.12), (2.15), (2.18)–(2.23) with initial data:

$$\varrho(r,0) = \varrho_0(r), \qquad 0 < r < R(0), \tag{3.1}$$

$$T(r,0) = T_0(r), \qquad 0 < r < R(0). \tag{3.2}$$

The following result holds.

THEOREM 3.1. *Let us assume that* $\varrho_0 \in C^{2+2\alpha}([0, R(0)])$, $0 < \alpha < \frac{1}{2}$. *Suppose also that for some* $\eta > 0$, $\varrho_0(r) \geq \eta$ *in the interval* $[0, R(0)]$. *Then there exists* $\delta = \delta(\|\varrho_0\|_{2+2\alpha})$ *such that the problem* (2.11), (2.12), (2.15), (2.18)–(2.23), (3.1), (3.2) *has a unique solution satisfying*

$$\varrho \in C^{2+2\alpha;1+\alpha}_{x;t}([0, R(t)] \times [0, \delta]),$$
$$R \in C^{3+\alpha}[0, \delta].$$

Proof. Using (2.11) we can eliminate T into (2.20) and deduce

$$\frac{p}{R_g}\frac{Ds}{Dt} = \kappa\Delta\left(\frac{p}{R_g\varrho}\right). \tag{3.3}$$

Taking into account (2.21) we can drop $p = p(t)$ from both sides of this equation, from which

$$\frac{Ds}{Dt} = \kappa\Delta\left(\frac{1}{\varrho}\right). \tag{3.4}$$

Plugging (2.12) in the left-hand side of this equation and using (2.19) we obtain

$$c_v\left\{\frac{p_t}{p} - \gamma\frac{D\varrho}{Dt}\right\} = \kappa\Delta\left(\frac{1}{\varrho}\right). \tag{3.5}$$

It readily follows from (2.20) that

$$\frac{1}{\varrho}\frac{D\varrho}{Dt} + \frac{1}{r^2}\frac{\partial}{\partial r}(r^2 v) = 0. \tag{3.6}$$

Plugging (3.6) into (3.5) we easily deduce

$$\frac{p_t}{p} = \frac{\kappa}{c_v}\Delta\left(\frac{1}{\varrho}\right) - \frac{\gamma}{r^2}\frac{\partial}{\partial r}(r^2 v). \tag{3.7}$$

Recalling that for radially symmetric functions $\Delta = \frac{1}{r^2}\frac{\partial}{\partial r}(r^2\frac{\partial}{\partial r})$, we multiply (3.7) by r^2 and integrate on the radial variable to obtain

$$\frac{p_t}{p}\frac{r^3}{3} = \frac{\kappa}{c_v}r^2\frac{\partial}{\partial r}\left(\frac{1}{\varrho}\right) - \gamma r^2 v + \alpha(t), \tag{3.8}$$

where $\alpha(t)$ is some suitable function depending only on t. We are interested in regular solutions at $r = 0$. Evaluating (3.8) at $r = 0$ it follows that $\alpha(t) = 0$. Equation (3.8) then becomes

$$\gamma v = \frac{\kappa}{c_v}\frac{\partial}{\partial r}\left(\frac{1}{\varrho}\right) - \frac{p_t}{p}\frac{r}{3}. \tag{3.9}$$

Eliminating v from (2.20) by means of (3.9) we deduce

$$\frac{\partial\varrho}{\partial t} + \frac{\kappa}{\gamma c_v}\frac{1}{r^2}\frac{\partial}{\partial r}\left(\varrho r^2\frac{\partial}{\partial r}\left(\frac{1}{\varrho}\right)\right) - \frac{p_t}{3\gamma p}\frac{1}{r^2}\frac{\partial}{\partial r}(\varrho r^3) = 0,$$

or, in an equivalent way,

$$\frac{\partial \varrho}{\partial t} = \frac{\kappa}{\gamma c_v}\Delta(\log(\varrho)) + \frac{p_t}{3\gamma p} r \frac{\partial \varrho}{\partial r} + \frac{p_t}{\gamma p}\varrho, \quad \text{in } B_{R(t)}(0). \tag{3.10}$$

On the other hand, the boundary condition (2.18) and the equation of state (2.11) yield

$$p(t) = R_g T_l \varrho(R(t), t). \tag{3.11}$$

Differentiating (3.11) we obtain

$$\frac{p_t}{p} = \left\{ \frac{\varrho_t}{\varrho} + \frac{\varrho_r}{\varrho}\dot{R}(t) \right\}. \tag{3.12}$$

Evaluating (3.9) at $r = R(t)$ and using the boundary condition (2.15) it follows that

$$\dot{R}(t) = -\frac{\kappa}{\gamma c_v} \frac{\varrho_r(R(t), t)}{(\varrho(R(t), t))^2} - \frac{R(t)}{3\gamma}\frac{p_t}{p}. \tag{3.13}$$

Summarizing, we have reduced the original problem to a free boundary problem for the density $\varrho(r, t)$ and the radius $R(t)$, namely (2.19), (3.10)–(3.13). To further simplify this problem we combine (2.19) and (3.12) to obtain

$$\frac{p_t}{p} = \frac{\Gamma(4\dot{R}\ddot{R} + R\dddot{R}) + p_{\infty,t} - \frac{2\sigma}{R^2}\dot{R}}{\Gamma(R\ddot{R} + \frac{3}{2}\dot{R}^2) + p_\infty + \frac{2\sigma}{R}}, \tag{3.14}$$

and substituting this formula into (3.13) we deduce a differential equation for the radius of the bubble:

$$\begin{aligned} \Gamma\dddot{R} = &-\frac{4\Gamma\dot{R}\ddot{R}}{R} - \frac{p_{\infty,t}}{R} + \frac{2\sigma\dot{R}}{R^3} \\ &-3\gamma\left(\frac{\kappa}{\gamma c_v}\frac{\varrho_r}{\varrho} + \dot{R}\right)\left[\Gamma\left(\frac{\ddot{R}}{R} + \frac{3}{2}\frac{(\dot{R})^2}{R^2}\right) + \frac{p_\infty}{R^2} + \frac{2\sigma}{R^3}\right]. \end{aligned} \tag{3.15}$$

On the other hand, eliminating $p(t)$ between (2.17) and (3.9) it follows that

$$\varrho(R(t), t) = \frac{1}{R_g T_l}\left[p_\infty + \frac{2\sigma}{R} + \Gamma\left(R\ddot{R} + \frac{3}{2}(\dot{R})^2\right)\right]. \tag{3.16}$$

We can obtain a solution of the free boundary problem (3.10), (3.14)–(3.16) by using a standard fixed point procedure in a way very similar to the one used in the analysis of the one-dimensional Stefan problem [Fr]. We transform the problem into a fixed boundary problem by means of the change of variables

$$r = R(t)y,$$

where y is the new space variable. Then problem (3.10), (3.14)–(3.16) becomes

$$\frac{\partial \varrho}{\partial t} = \frac{\kappa}{c_v}\Delta_y(\log(\varrho)) + f_1\left(y, t, \varrho(y, t), \quad \frac{\partial \varrho}{\partial y}(y, t), \varrho(1, t), R, \dot{R}, \ddot{R}, \dddot{R}\right) \quad \text{in } B_1(0), \tag{3.17}$$

$$\varrho(1,t) = f_2(R, \dot{R}, \ddot{R}, t), \tag{3.18}$$

$$\dddot{R} = f_3\left(t, R, \dot{R}, \ddot{R}, \varrho(1,t), \frac{\partial \varrho}{\partial y}(1,t)\right), \tag{3.19}$$

where the functions f_1, f_2, f_3 are analytic functions on all their variables for $R \neq 0$, and $\varrho(1,t) \neq 0$. This problem is equivalent to the original free boundary problem for smooth solutions.

The local solution of (3.17)–(3.19) with initial condition (3.1) as well as the initial conditions $R(0) = R_0, \dot{R}(0) = \dot{R}_0$ can be obtained by means of a standard fixed point argument on $R(t)$. More precisely, we introduce a functional space:

$$X = \{\bar{R}(t) : \bar{R} \in C^{3+\alpha}[0,\delta], \bar{R}(0) = R_0, \quad \dot{\bar{R}}(0) = \dot{R}_0, [\bar{R}]_{3+\alpha} < \infty\},$$

where $\delta > 0$ is some small number that will be chosen later. Given $\bar{R} \in X$ we solve (3.15), (3.16) with R replaced by $\bar{R}$ with initial data $\varrho_0(y,0) = \varrho_0(R_0 y)$. Taking into account that $\varrho_0 \geq \eta > 0$ for $0 \leq r \leq R_0$ we can apply classical regularity theory for quasi-linear parabolic equations in order to deduce the existence and uniqueness of a smooth solution $\varrho(y,t)$ for $0 \leq t \leq \delta \leq \delta_0(\eta, \|\varrho_0\|_{L^\infty(B_1(0))})$. Moreover, by standard estimates, it follows that

$$[\varrho]_{C^{2+2\alpha;1+\alpha}_{y;t}(B_1(0)\times[0,\delta])} \leq C(\|\varrho_0\|_{L^\infty(B_1(0))}, \eta),$$

where $0 < \alpha < \frac{1}{2}$. It then follows by classical embedding theory that

$$\left[\frac{\partial \varrho}{\partial y}(1,\cdot)\right]_{C^{\frac{1}{2}+\alpha}_t([0,\delta])} \leq C(\|\varrho_0\|_{L^\infty(B_1(0))}, \eta).$$

Then, given $\bar{R} \in X$ we can define the operators:

$$T_1[\bar{R}] \equiv \varrho(1,t),$$
$$T_2[\bar{R}] \equiv \frac{\partial \varrho}{\partial y}(1,t),$$

where $\varrho(y,t)$ is the corresponding solution of (3.17), (3.18) that we have just obtained. The pair $(\bar{R}, \varrho)$, where $\bar{R} \in X$, solves the problem (3.17)–(3.19) if and only if $\bar{R}$ solves the equation

$$\dddot{R} = f_3(t, R, \dot{R}, \ddot{R}, T_1[R], T_2[R]), \qquad 0 \leq t \leq \delta, \tag{3.20}$$

where $R(0) = R_0, \dot{R}(0) = \dot{R}_0$, and $\ddot{R}$ is obtained from (3.16) as

$$\ddot{R}(0) = \frac{1}{R_0}\left\{\frac{1}{\Gamma}\left(R_g T_l \varrho_0(y=1) - p_\infty(0) - \frac{2\sigma}{R_0}\right) - \frac{3}{2}(\dot{R}_0)^2\right\}.$$

Taking into account the analyticity of f_1, f_2 and classical continuous dependence of the solutions of (3.17) on the source term R it follows that for $\bar{R}_1, \bar{R}_2 \in X$

$$[T_1[\bar{R}_1] - T_1[\bar{R}_2]]_{C^{1+\alpha}[0,\delta]} + [T_2[\bar{R}_1] - T_2[\bar{R}_2]]_{C^{\frac{1}{2}+\alpha}[0,\delta]} \leq C[\bar{R}_1 - \bar{R}_2]_{C^{3+\alpha}[0,\delta]}. \tag{3.21}$$

We can write (3.20) as

$$
\begin{aligned}
&R(t) = R_0 + \dot{R}_0 t + \ddot{R}_0 \frac{t^2}{2} \\
&+ \int_0^t \int_0^{s_1} \int_0^{s_2} f_3(s_3, R(s_3), \dot{R}(s_3), \ddot{R}(s_3), T_1[R](s_3), T_2[R](s_3)) ds_3 ds_2 ds_1.
\end{aligned} \tag{3.22}
$$

Using (3.21) it easily follows that the operator on the right-hand side of (3.22) is contractive on the space X if δ is small enough. The proof of Theorem 3.1 then follows by means of a classical fixed point argument. □

4. Stability of the steady state. The goal of this section is to prove that the steady state of (2.11), (2.12), (2.15), (2.18)–(2.23) is stable under small perturbations.

4.1. Description of the steady state. We have reduced the original problem with $\Gamma > 0$ to the free boundary problem (3.10), (3.14)–(3.16). We now proceed to describe the steady state of this problem. To this end, we assume without loss of generality that the forcing term $p_\infty(t)$ is constant:

$$p_\infty \equiv 1.$$

We are interested in steady state solutions. Taking into account (3.10), (3.13), we deduce that the steady state satisfies

$$\Delta(\log(\varrho)) = 0 \qquad \text{in } B_{R_s}(0),$$

$$\frac{\partial \varrho}{\partial r}(R_s) = 0,$$

where R_s denotes the stationary radius. Taking into account that ϱ depends only on r it follows that

$$\varrho(r) = \varrho(R_s) \equiv \varrho_0 \qquad \text{for} \qquad 0 \le r \le R_s. \tag{4.1}$$

On the other hand, (4.1), (2.11), (2.13), and the boundary condition (2.16) imply

$$T(r) = T_l \qquad \text{for} \qquad 0 \le r \le R_s, \tag{4.2}$$

from which by (2.11), (2.19) it follows that

$$R_g \varrho_0 T_l = p_\infty + \frac{2\sigma}{R_s} \equiv p_0, \tag{4.3}$$

which fixes the radius R_s as a function of ϱ_0. Summarizing, for each value of ϱ_0 we have one value of R_s prescribed by (4.3) that corresponds to the stationary radius.

4.2. Helmholtz free energy. In this section we recall some classical thermodynamic expressions that will be useful in our forthcoming analysis of the stability of steady states. Notice that the gas contained in the bubble is doing mechanical work against the liquid in a process that takes place at a constant boundary temperature. It is well known [LL] that for such processes the Helmholtz free energy provides a suitable Lyapunov function for the evolution of the system. The main goal of this section is to write the evolution law for the Helmholtz free energy in our particular setting. Given a function $\varphi(x,t)$ the following formula holds:

$$\frac{d}{dt}\left(\int_{B_{R(t)}} \varrho \varphi d^3x\right) = \int_{B_{R(t)}} \varrho \frac{D\varphi}{Dt} d^3x, \tag{4.4}$$

where $\partial B_{R(t)}$ moves with the fluid. Although (4.4) is rather standard, for the reader's convenience, we recall its derivation in the radially symmetric case is given in Appendix B at the end of the paper.

On the other hand, the following well-known thermodynamic identity is satisfied:

$$du = Tds - pd\left(\frac{1}{\varrho}\right), \tag{4.5}$$

where u is the internal energy by unit of mass.

In the case of an ideal gas, the entropy per unit of gas is given by (2.12) which combined with (4.5) implies the well-known thermodynamic identity:

$$u = c_v T, \tag{4.6}$$

where we have used $\gamma = \frac{c_p}{c_v}$ as well as $c_p - c_v = R_g$.

Notice that (4.5) implies

$$\varrho \frac{Du}{Dt} = \varrho T \frac{Ds}{Dt} - \varrho p \frac{D(\frac{1}{\varrho})}{Dt} = \varrho T \frac{Ds}{Dt} + \frac{p}{\varrho}\frac{D\varrho}{Dt}. \tag{4.7}$$

Using (2.10), (4.7), as well as the continuity equation (2.8) written in the form

$$\frac{D\varrho}{Dt} + \varrho \ \mathrm{div}(\overrightarrow{v}) = 0,$$

and it readily follows that

$$\varrho \frac{Du}{Dt} = \nabla(\kappa \nabla T) - p \ \mathrm{div}(\overrightarrow{v}). \tag{4.8}$$

The Helmholtz free energy for the bubble is given by the thermodynamic expression:

$$F = U - T_l S,$$

where U and S are the total internal energy and the total entropy of the gas, respectively, and T_l is the temperature of the heat bath (in our case the liquid).

In our case the Helmholtz free energy of the gas inside the bubble is given by the expression

$$F = \int_{B_{R(t)}} \varrho u d^3x - T_l \int_{B_{R(t)}} \varrho s d^3x. \tag{4.9}$$

Taking into account (4.4), and using (2.10), (2.13), (2.15), and (4.8), it follows that

$$\begin{aligned}
\frac{dF}{dt} &= \int_{B_{R(t)}} \nabla(\kappa \nabla T) d^3x - p(t) \int_{B_{R(t)}} \mathrm{div}(\overrightarrow{v}) d^3x - \int_{B_{R(t)}} \frac{T_l}{T} \nabla(\kappa \nabla T) d^3x \\
&= \int_{\partial B_{R(t)}} \kappa \nabla T d\overrightarrow{S} - p(t) \int_{\partial B_{R(t)}} v d\overrightarrow{S} \\
&\quad - \int_{\partial B_{R(t)}} \kappa \nabla T d\overrightarrow{S} - T_l \int_{B_{R(t)}} \kappa \frac{(\nabla T)^2}{T^2} d^3x \\
&= -\kappa T_l \int_{B_{R(t)}} \frac{(\nabla T)^2}{T^2} d^3x - p(t) v(R(t), t) \int_{\partial B_{R(t)}} dS.
\end{aligned} \tag{4.10}$$

The first term on the right-hand side of (4.11) corresponds to the dissipation of energy due to thermal conductivity. The second is the total amount of work made for the bubble in its expansion. Writing

$$\frac{dW}{dt} = -p(t)v(R(t),t)\int_{\partial B_{R(t)}} dS = -4\pi p(t)(R(t))^2\dot{R}(t),$$

we obtain from (4.11) the well-known thermodynamic formula:

$$\frac{dF}{dt} \leq \frac{dW}{dt}.$$

We now use the fact that the amount of work made by the gas against the liquid is spent in the variation of the interfacial energy of the liquid-gas boundary as well as in the change of mechanical energy of the liquid. The kinetic energy of the liquid is

$$K_l = \frac{1}{2}\int_{R(t)}^{+\infty} \varrho_l 4\pi r^2 (v_r(r,t))^2 dr, \tag{4.11}$$

where $v_r(r,t)$ has to be computed using (2.2), (2.6) implying

$$K_l = 2\pi\varrho_l (R(t))^3(\dot{R}(t))^2. \tag{4.12}$$

On the other hand, the energy of the liquid-gas interface is given by

$$U_{gl} = 4\pi\sigma(R(t))^2. \tag{4.13}$$

We compute

$$\frac{d}{dt}(U_{gl}+K_l) = 6\pi\varrho_l R^2\dot{R}^3 + 4\pi\varrho_l R^3\dot{R}\ddot{R} + 8\pi\sigma R\dot{R}.$$

Using (2.7) to compute $\dot{R}\ddot{R}$ we deduce that

$$\frac{d}{dt}(U_{gl}+K_l) = 6\pi\varrho_l R^2\dot{R}^3 - 6\varrho_l R^2\dot{R}^3 + 4\pi R^2(p_l(t)-p_\infty(t))\dot{R} + 8\pi\sigma R\dot{R}.$$

With the help of the boundary condition (2.14), it then follows that

$$\frac{d}{dt}(U_{gl}+K_l) = -4\pi p_\infty(t)R^2\dot{R} + 4\pi R^2 p(t)\dot{R}. \tag{4.14}$$

Adding (4.11) and (4.14) we arrive at

$$\frac{d}{dt}(F+U_{gl}+K_l) = -4\pi p_\infty(t)R^2\dot{R} - \kappa T_l\int_{B_{R(t)}}\frac{(\nabla T)^2}{T^2}d^3x. \tag{4.15}$$

The first term on the right-hand side of (4.15) is the work made by the external sound field. The second term is the energy dissipation due to the termal conductivity of the gas.

In the particular case in which we take $\alpha = 0$ in (2.21), (4.15) may be written as

$$\frac{d}{dt}\left(F+U_{gl}+K_l+\frac{4\pi}{3}R^3\right) = -\kappa T_l\int_{B_{R(t)}}\frac{(\nabla T)^2}{T^2}d^3x. \tag{4.16}$$

Equation (4.16) is just a version of the second law of thermodynamics particularized to our actual problem. Notice that (4.16) implies that the steady state is characterized for a uniform distribution of temperature T inside the bubble. Moreover, at those points the functional $F+U_{gl}+K_l+\frac{4\pi}{3}R^3$ reaches an extremal point under our current assumptions. Notice that since $\nabla T = 0$ inside the bubble, the boundary condition (2.16) implies $T = T_l$. Taking into account that in our model the pressure is uniform and using the equation of state (2.11), it follows that ϱ is uniform. We can then write (4.9) as

$$F = \frac{4\pi R^3}{3}\varrho u - \frac{4\pi R^3}{3}\varrho s T_l,$$

whence the steady state solutions correspond to the extremal points of the function

$$\Phi \equiv \frac{4\pi R^3}{3}\varrho u - \frac{4\pi R^3}{3}\varrho s T_l + 4\pi\sigma R^2 + \frac{4\pi}{3}R^3, \tag{4.17}$$

where (p, s, u) are given, respectively, by (2.11), (2.12), and (4.6) and where the conservation of mass inside the bubble implies

$$\frac{4\pi R^3}{3}\varrho = M_0. \tag{4.18}$$

Using (2.11), (2.12) and (4.6), (4.18) we can rewrite (4.17) as

$$\Phi = c_v T_l M_0 - c_v M_0 T_l \log(T_l) + c_v(\gamma-1)T_l M_0 \log(\varrho) + 4\pi\sigma R^2 + \frac{4\pi}{3}R^3. \tag{4.19}$$

We eliminate ϱ from (4.19) using (4.18). It then follows (using also the classical formula) that

$$c_v(\gamma-1) = c_p - c_v = R_g,$$

$$\Phi = c_v T_l M_0 - c_v M_0 T_l \log(T_l) + R_g T_l M_0 \log\left(\frac{3M_0}{2\pi}\right) - 3R_g T_l M_0 \log(R) + 4\pi\sigma R^2 + \frac{4\pi}{3}R^3,$$

and computing the minimum of the function with respect to R we readily deduce that the equilibrium $R = R_s$ satisfies

$$p_0 \equiv R_g \varrho_0 T_l = 1 + \frac{2\sigma}{R_s},$$

where ϱ_0 is the density inside the gas. As it should be expected this formula coincides with (4.3).

4.3. Linearized problem near the steady state. Our next goal is to study the linearization of the problem (2.11), (2.12), (2.15), (2.18)–(2.23) near the steady state, where we take $\alpha = 0$ in (2.23). We consider the evolution equation written in the form (3.10), (3.14)–(3.16). We then set

$$p = p_0 + g, \tag{4.20a}$$

$$\varrho = \varrho_0 + \zeta, \tag{4.20b}$$

$$R = R_s + \lambda. \tag{4.20c}$$

Formally linearizing in (3.8), (3.12), (3.13), (3.14) we then obtain the following system of equations:

$$\frac{\partial \zeta}{\partial t} = \frac{\kappa}{c_v T_l}\Delta(\zeta) + \frac{1}{R_g} T_l \left(\Gamma R_s \dddot{\lambda} - \frac{2\sigma}{R_s^2}\dot{\lambda} \right), \tag{4.21}$$

$$\Gamma \dddot{\lambda} = -3\frac{p_0}{R_s^2}\left(\dot{\lambda} + \frac{\kappa}{c_v}\frac{\zeta_r(R_s,t)}{\varrho_0^2} \right) + \frac{2\sigma}{R_s^3}\dot{\lambda}, \tag{4.22}$$

$$\zeta(R_s,t) = \frac{1}{R_g T_l}\left(-\frac{2\sigma}{R_s^2}\lambda + \Gamma R_s \ddot{\lambda} \right), \tag{4.23}$$

where by assumption $\lambda(0), \dot{\lambda}(0), \zeta(r,0)$ are given and $\ddot{\lambda}(0)$ can be then obtained from (4.23).

On the other hand, if we keep in (4.15) just the quadratic terms on g, ζ, λ we obtain a formula that is satisfied for the linearized system (4.21)–(4.23). We could do this more precisely by writing (4.20) as

$$p = p_0 + \varepsilon g,$$

$$\varrho = \varrho_0 + \epsilon\zeta,$$

$$R = R_s + \varepsilon\lambda.$$

Then taking the formal limit $\varepsilon = 0$ in the linear equation and in the energy dissipation formula (4.16) written in new variables it would follow the energy formula for the linearized problem that contains just quadratic terms. More precisely, we can write

$$K_l = 2\pi\varrho_l R_s^3(\dot{\lambda})^2 + \text{cubic terms},$$

$$U_{gl} = 4\pi\sigma R_s^2 + 8\pi\sigma R_s\lambda + 4\pi\sigma\lambda^2,$$

$$\frac{4\pi}{3}R^3 = \frac{4\pi}{3}R_s^3 + 4\pi R_s^2\lambda + 4\pi R_s\lambda^2 + \text{cubic terms}.$$

We approximate F as follows:

$$\begin{aligned} F &= c_v\int_{B_R}\varrho T d^3x - c_v T_l\int_{B_R}\varrho\log(p)d^3x + c_v\gamma T_l\int_{B_R}\varrho\log(\varrho)d^3x \\ &= \frac{c_v p}{R_g}\int_{B_R} d^3x - c_v T_l M_0\log(p) + c_v\gamma T_l\int_{BR}\varrho\log(\varrho)d^3x \\ &= \frac{4\pi c_v p}{3R_g}R^3 - c_v T_l M_0\log(p) + c_v\gamma T_l\int_{B_R}\varrho\log(\varrho)d^3x, \end{aligned}$$

where

$$M_0 = \int_{B_R} \varrho d^3x$$

is the total mass of gas, and we consider perturbations of the gas that keep the pressure p homogeneous throughout volume.

We write

$$\begin{aligned}\int_{B_R} \varrho \log(\varrho) d^3x &= \int_{B_R} \varrho \log(\varrho+\zeta) d^3x \\ &= \int_{B_R} \varrho \left\{ \log(\varrho_0) + \frac{\zeta}{\varrho_0} - \frac{\zeta^2}{2\varrho_0^2} \right\} d^3x + \text{cubic terms} \\ &= \log(\varrho_0) M_0 + \int_{B_R} \frac{\varrho}{\varrho_0} \zeta d^3x - \frac{1}{2\varrho_0} \int_{B_{R_0}} \zeta^2 d^3x + \text{cubic terms} \\ &= M_0 \log(\varrho_0) + \int_{B_R} \zeta d^3x + \frac{1}{2\varrho_0} \int_{B_{R_0}} \zeta^2 d^3x + \text{cubic terms}.\end{aligned}$$

Using the fact that

$$M_0 = \int_{B_R} (\varrho_0 + \zeta) d^3x, \tag{4.24}$$

it follows that

$$\int_{B_R} \zeta d^3x = M_0 - \int_{B_R} \varrho_0 d^3x = M_0 - \varrho_0 \frac{4\pi R^3}{3},$$

from which

$$\begin{aligned}F &= \frac{4\pi c_v p R^3}{3R_g} - c_v T_l M_0 \log(p) + c_v \gamma T_l M_0 \log(\varrho_0) \\ &\quad + c_v \gamma T_l M_0 - c_v \gamma T_l \varrho_0 \frac{4\pi R^3}{3} + \frac{c_v \gamma T_l}{2\varrho_0} \int_{B_{R_0}} \zeta^2 d^3x + \text{cubic terms} \\ &= F_0 + \left(\frac{4\pi c_v p_0 R_s^2 \lambda}{R_g} - c_v \gamma T_l 4\pi R_s^2 \varrho_0 \lambda \right) \\ &\quad + \left[\frac{4\pi c_v p_0 R_s}{R_g} \lambda^2 + \frac{4\pi c_v R_s^2}{R_g} g\lambda + \frac{c_v T_l M_0}{2p_0^2} g^2 \right. \\ &\quad \left. -4\pi c_v \gamma T_l \varrho_0 R_s \lambda^2 + \frac{c_v \gamma T_l}{2\varrho_0} \int_{B_{R_s}} \zeta^2 d^3x \right] \\ &\quad + \text{cubic terms},\end{aligned}$$

where

$$F_0 = \frac{4\pi c_v p_0 R_s^3}{3R_g} - c_v T_l M_0 \log(p_0) + c_v \gamma T_l M_0 \log(\varrho_0) + c_v \gamma T_l M_0 - \gamma c_v T_l \varrho_0 \frac{4\pi}{3} R_s^3.$$

From this we derive

$$\begin{aligned}F &= F_0 - 4\pi c_v R_s^2 (\gamma - 1) T_l \varrho_0 \lambda \\ &\quad + \left[4\pi c_v \varrho_0 T_l R_s (1-\gamma) \lambda^2 + \frac{4\pi c_v R_s^2}{R_g} g\lambda + \frac{c_v T_l M_0}{2p_0^2} g^2 + \frac{c_v \gamma T_l}{2\varrho_0} \int_{B_{R_s}} \zeta^2 d^3x \right] \\ &\quad + \text{cubic terms}.\end{aligned}$$

Summarizing to quadratic order and using that $p_0 = 1 + \frac{2\sigma}{R_s}$ we deduce from (4.16) that the solution of (4.21)–(4.23) satisfies

$$\begin{aligned} &\frac{d}{dt}\left[4\pi\sigma\lambda^2 + \tfrac{4\pi c_v R_s^2}{R_g}g\lambda + \tfrac{c_v T_l M_0}{2p_0^2}g^2 + \tfrac{c_v\gamma T_l}{2\varrho_0}\int_{B_{R_s}}\zeta^2 d^3x + 2\pi\varrho_l R_s^3(\lambda_t)^2\right] \\ &= -\tfrac{\kappa T_l}{\varrho_0^2}\int_{B_{R_s}}|\nabla\zeta|^2 d^3x. \end{aligned} \tag{4.25}$$

Notice that (4.24) implies to linear order

$$\lambda = -\frac{1}{4\pi\varrho_0 R_s^2}\int_{B_{R_s}}\zeta d^3x, \tag{4.26}$$

which in turn yields

$$\begin{aligned} &4\pi\sigma\lambda^2 + \frac{4\pi c_v R_s^2}{R_g}g\lambda + \frac{c_v T_l M_0}{2p_0^2}g^2 + \frac{c_v T_l\gamma}{2\varrho_0^2}\int_{B_{R_s}}\zeta^2 d^3x \\ &= \frac{\sigma}{4\pi\varrho_0^2 R_s^4}\left(\int_{B_{R_s}}\zeta d^3x\right)^2 - \frac{c_v}{\varrho_0 R_g}g\int_{B_{R_s}}\zeta d^3x + \frac{c_v T_l M_0}{2p_0^2}g^2 + \frac{c_v\gamma T_l}{2\varrho_0}\int_{B_{R_s}}\zeta^2 d^3x \\ &= \frac{c_v T_l M_0}{2}\left(\frac{g}{p_0} - \frac{1}{M_0}\int_{B_{R_s}}\zeta d^3x\right)^2 + \frac{c_v\gamma T_l}{2\varrho_0}\int_{B_{R_s}}\zeta^2 d^3x \\ &- \left(\frac{\sigma}{4\pi\varrho_0^2 R_s^4} + \frac{c_v T_l}{2M_0}\right)\left(\int_{B_{R_s}}\zeta d^3x\right)^2 \geq \frac{c_v T_l M_0}{2}\left(\frac{g}{p_0} - \frac{1}{M_0}\int_{B_{R_s}}\zeta d^3x\right)^2 \\ &+ \left[\frac{c_v\gamma T_l}{2\varrho_0} - \left(\frac{\sigma}{4\pi\varrho_0^2 R_s^4} + \frac{c_v T_l}{2M_0}\right)\frac{4\pi R_s^3}{3}\right]\int_{B_{R_s}}\zeta^2 d^3x \\ &= \frac{c_v T_l M_0}{2}\left(\frac{g}{p_0} - \frac{1}{M_0}\int_{B_{R_s}}\zeta d^3x\right)^2 \\ &+ \frac{1}{\varrho_0}\left(\frac{c_v(\gamma-1)T_l}{2} - \frac{R_g T_l}{6} + \frac{1}{6\varrho_0}\right)\int_{B_{R_s}}\zeta^2 d^3x \\ &= \frac{c_v T_l M_0}{2}\left(\frac{g}{p_0} - \frac{1}{M_0}\int_{B_{R_s}}\zeta d^3x\right)^2 + \left(\frac{R_g T_l}{3} + \frac{1}{6\varrho_0}\right)\int_{B_{R_s}}\zeta^2 d^3x. \end{aligned} \tag{4.27}$$

Using the energy estimate (4.16) we can prove the following stability result.

THEOREM 4.1. *Let us assume that $\varrho(r,t)$, $R(t)$ is a solution of the free boundary problem* (3.10), (3.14)–(3.16) *with $p_\infty = 1$ and with initial data $\varrho(r,0)$, $R(0)$ satisfying*

$$[\varrho(\cdot,0) - \varrho_0]_{C_r^{2+2\alpha}} \leq \eta_0, \tag{4.28}$$

$$|R(0) - R_s| \leq \eta_0, \tag{4.29}$$

where $\eta_0 > 0$ is small enough and η_0, R_s are the values of the density and the radius corresponding to a steady state (cf. (4.3)*). Then the solution $(\varrho(r,t), R(t))$ is globally defined and satisfies the estimates*

$$[\varrho(\cdot,,t) - \varrho_0]_{C_r^{2+2\alpha}} \leq \varepsilon_0, \tag{4.30}$$

$$|R(t) - R_0| \leq \varepsilon_0 \tag{4.31}$$

for any $t > 0$, where ε_0 is a small number that can be made arbitrarily small if η_0 in (4.28), (4.29) *is small enough.*

Theorem 4.1 shows that steady states are Lyapunov stable. The proof of Theorem 4.1 is a standard use of the classical Lyapunov method. A key step in the proof of Theorem 4.1 is the following lemma.

LEMMA 4.2. *Let us assume that there exists a constant $\nu > 1$ such that*

$$\frac{1}{\nu} \leq \varrho(r) \leq \nu, \tag{4.32}$$

$$\frac{1}{\nu} \leq R \leq \nu, \tag{4.33}$$

$$\int_{B_R} \varrho(r) d^3x = M_0. \tag{4.34}$$

Then there exist constants $\Theta > 0$, $\delta_0 > 0$ depending only on M_0 in (4.18), *T_l in* (2.18), *and ν such that the function $Q = F + U_{gl} + \frac{4\pi}{3} R^3$ satisfies*

$$Q - Q_0 \geq \Theta \left(\int_{B_R} (\varrho(r) - \varrho_0)^2 d^3x \right), \tag{4.35}$$

provided that $|\varrho - \varrho_0| \leq \delta_0$, and where Q_0 is evaluated at the steady state (ϱ_0, R_s).

Proof. The proof of Lemma 4.2 consists essentially in retracing the steps in the proof of (4.27) and keeping track of the cubic terms that were neglected there. In this way we obtain the following estimate:

$$\begin{aligned} Q - Q_0 \geq{}& \frac{c_v T_l M_0}{2} \left(\frac{p - p_0}{p_0} - \frac{1}{M_0} \int_{B_R} (\varrho - \varrho_0) d^3x \right)^2 \\ &+ \left(\frac{R_g T_l}{3} + \frac{1}{6\varrho_0} \right) \left(\int_{B_R} (\varrho - \varrho_0)^2 d^3x \right) \\ &- C \left(\int_{B_R} |\varrho - \varrho_0|^3 d^3x + |p - p_0|^3 + |R - R_0|^3 \right), \end{aligned} \tag{4.36}$$

where the constant C depends only on ν, M_0, T_l. Notice that the integrals in (4.36) are computed in B_R instead of in B_{R_s} as in (4.27). Using (4.24) we obtain the following nonlinear version of (4.26):

$$\frac{4\pi\varrho_0}{3}(R^3 - R_s^3) = -\int_{B_R} (\varrho - \varrho_0) d^3x. \tag{4.37}$$

Taking into account (4.33) it readily follows that

$$|R - R_s| \leq C \left(\int_{B_R} |\varrho - \varrho_0|^2 d^3x \right)^{\frac{1}{2}}, \tag{4.38}$$

where C depends only on ν, M_0, T_l. On the other hand,

$$|p - p_0| \leq C \left\{ \left| \frac{p - p_0}{p_0} - \frac{1}{M_0} \int_{B_R} (\varrho - \varrho_0) d^3x \right| + \left(\int_{B_R} (\varrho - \varrho_0)^2 d^3x \right)^{\frac{1}{2}} \right\}. \tag{4.39}$$

Using (4.38), (4.39) and taking into account that $|p-p_0| = R_g T_l |\varrho - \varrho_0|$ at $r = R$, we readily obtain that for $|\varrho - \varrho_0| \leq \delta_0$ and δ_0 small enough (4.35) holds true. □

Proof of Theorem 4.1. It is easily checked using (4.4) that the total mass of gas $\int_{B_R} \varrho d^3x$ is a constant of the motion. We rewrite (4.16) as

$$\frac{d}{dt}(Q + K_l) = -KT_l \int_{B_R} \frac{(\nabla T)^2}{T^2} d^3x, \tag{4.40}$$

where Q is as in Lemma 4.2. We then obtain the following estimate:

$$Q(t) - Q_0 \leq K_l(0) + (Q(0) - Q_0). \tag{4.41}$$

Under the assumptions of Theorem 4.1, we can take $|K_l(0)|$ and $|Q(0) - Q_0|$ smaller than some number $\varepsilon_1 > 0$ if η_0 is chosen small enough. On the other hand, using Lemma 4.2, combined with (4.41) we obtain

$$\Theta\left(\int_{B_R} (\varrho(r,t) - \varrho_0)^2 d^3x\right) \leq \varepsilon_1. \tag{4.42}$$

Let us assume that

$$[\varrho(\cdot,t) - \varrho_0]^{2+2\alpha}_{C_r} \leq \nu \tag{4.43}$$

for some $\nu > 1$. Certainly this estimate holds for short times, and as far as this bound and (4.32) are satisfied we can apply Theorem 3.1 in order to define the solution $(\varrho(r,t), R(t))$ for a larger time. Our goal is to show (4.43) globally in time. To this end, we argue as follows. Using classical interpolation results we can obtain from (4.42), (4.43) the estimate

$$\left|\frac{\partial \varrho}{\partial r}\right| + |\varrho - \varrho_0| \leq \varepsilon_2, \tag{4.44}$$

where ε_2 might be made small if ε is so.

Notice that (4.44) implies, if ε_2 is small enough, that (4.32) holds as far as (4.43) is satisfied. Moreover, (4.38) and (4.44) imply (4.33). We can then combine (3.15) and the estimate for $\frac{\partial \varrho}{\partial r}$ in (4.44) to show that $|\dddot{R}|$ is uniformly small. The proof can then be concluded by means of a bootstrap argument. More precisely, the regularity of $\ddot{R}$ as well as (3.10), (3.14), (3.16) implies estimates for $[\frac{\partial \varrho}{\partial r}(r=1,\cdot)]_\delta$, $0 < \delta < \frac{1}{4}$, and this provides similar estimates for $[\dddot{R}(\cdot)]_\delta$. Iterating the procedure, we obtain bounds for $[\dddot{R}(\cdot)]_\alpha$ for any $\alpha < \frac{1}{2}$ as well as analogous estimates for ϱ. This yields (4.43) and concludes the proof of Theorem 4.1 □

5. The limit case $\Gamma = 0$. We now consider the problem (2.11), (2.12), (2.15), (2.18)–(2.23) in the case in which we formally make $\Gamma = 0$, which is rather natural since as indicated before, in nondimensional units $\Gamma \sim \frac{1}{400}$. Recalling (2.15), (2.19), (3.7), (3.8), (3.9) it follows that the problem can be written as the following free boundary problem:

$$\frac{\partial \varrho}{\partial t} = \frac{\kappa}{\gamma c_v} \Delta(\log(\varrho)) + \frac{p_t}{3\gamma p} r \frac{\partial \varrho}{\partial r} + \frac{p_t}{\gamma p} \varrho, \tag{5.1}$$

$$p(t) = R_g T_l \varrho(R(t), t), \tag{5.2}$$

$$\dot{R}(t) = \frac{\kappa}{\gamma c_v}\frac{\partial}{\partial r}\left(\frac{1}{\varrho}\right)\Big|_{r=R(t)} - \frac{p_t}{p}\frac{R}{3\gamma}, \tag{5.3}$$

$$p(t) = p_\infty(t) + \frac{2\sigma}{R(t)}. \tag{5.4}$$

Problem (5.1)–(5.4) with suitable initial data is a standard free boundary problem very similar to the classical one-dimensional Stefan problem. Local existence and uniqueness of Hölder solutions of (5.1)–(5.4) can be obtained with an argument similar to that used in the proof of Theorem 3.1. The following result holds.

THEOREM 5.1. *Assume that* $\varrho_0 \geq \eta > 0$, $\varrho_0 \in C^{2,\alpha}([0, R(0)])$, $\frac{\partial \varrho_0}{\partial r}(0) = 0$. *Then for* $\delta > 0$ *small enough there exists a unique solution of* (5.1)–(5.4) *in the class of function* $R \in C^{1+\alpha}([0,\delta])$, $0 < \alpha < \frac{1}{2}$, $\varrho \in C^{1+\frac{\alpha}{2},2+\alpha}([0,\delta] \times [0, R(t)])$.

The proof of Theorem 5.1 can be obtained exactly as the analogous result for the one-dimensional Stefan problem [Fr]. It turns out that solutions of (5.1)–(5.4) are global in time if the external pressure p_∞ remains always positive. More precisely, we have the following.

THEOREM 5.2. *If* $p_\infty > 0$, *any solution of* (5.1)–(5.4) *with initial data* $\varphi(r,0) = \varphi_0(r) > 0$ *is global, i.e., for any* $T > 0$ *there holds*

$$|\dot{R}(t)| \leq C(T), \quad 0 \leq t \leq T,$$

for some suitable function $C(T)$.

Proof. We define a new variable as

$$\varphi(r,t) = \frac{\varrho(r,t)}{R_g T_l p(t)}, \tag{5.5}$$

that transforms (5.1)–(5.4) into

$$\varphi_t = \frac{A}{p(t)}\Delta(\log(\varphi)) + \frac{p_t}{3\gamma p} r\varphi_r + \left(\frac{1}{\gamma} - 1\right)\frac{p_t}{p}\varphi, \qquad A = \frac{\kappa}{\gamma c_v R_g T_l}, \tag{5.6}$$

$$\varphi(R(t),t) = 1, \tag{5.7}$$

$$\dot{R}(t) = -\frac{A}{p(t)}\varphi_r(R,t) - \frac{p_t}{3\gamma p}R. \tag{5.8}$$

If we introduce the change of variables $\xi = \frac{r}{R(t)}$, (5.6) becomes

$$\varphi_t = \frac{A}{pR^2}\Delta_\xi(\log(\varphi)) + \left(\frac{p_t}{3\gamma p} + \frac{R_t}{R}\right)\xi\varphi_\xi + \left(\frac{1}{\gamma} - 1\right)\frac{p_t}{p}\varphi. \tag{5.9}$$

Notice that

$$\frac{R_t}{R} + \frac{p_t}{3\gamma p} = \frac{R_t}{R} + \frac{1}{3\gamma p}\left(p_{\infty,t} - \frac{2\sigma R_t}{R^2}\right) = \frac{R_t}{R}\left(1 - \frac{2\sigma}{R}\frac{1}{3\gamma p}\right) + \frac{p_{\infty,t}}{3\gamma p}. \tag{5.10}$$

On the other hand, (5.8) implies

$$R_t = -\frac{A}{p(t)}\frac{1}{R}\varphi_\xi(R,t) - \frac{R}{3\gamma p}\left(p_{\infty,t} - \frac{2\sigma R_t}{R^2}\right),$$

whence, using (5.10) it follows that

$$\frac{R_t}{R}+\frac{p_t}{3\gamma p}=-\frac{A}{pR^2}\varphi_\xi(1,t)-\frac{1}{3\gamma p}p_{\infty,t}+\frac{p_{\infty,t}}{3\gamma p}=-\frac{A}{pR^2}\varphi_\xi(1,t).$$

We then write (5.9) as

$$\varphi_t=\frac{A}{pR^2}\Delta_\xi(\log(\varphi))-\frac{A}{pR^2}\varphi_\xi(1,t)\xi\varphi_\xi+\left(\frac{1}{\gamma}-1\right)\frac{p_t}{p}\varphi, \tag{5.11}$$

which we have to complement with the boundary condition

$$\varphi(1,t)=1 \tag{5.12}$$

as well as (5.8) which we rewrite the new variables as

$$R_t=\left(1-\frac{2\sigma}{R}\frac{1}{3\gamma p}\right)^{-1}\left[-\frac{A}{pR}\varphi_\xi(1,t)-\frac{R}{3\gamma p}p_{\infty,t}\right]. \tag{5.13}$$

Notice that since $\gamma>1$ our assumptions on $p_\infty(t)$ imply that $(1-\frac{2\sigma}{R}\frac{1}{3\gamma p})^{-1}$ is a smooth function as far as the solution of (5.1)–(5.4) is defined. Before proving Theorem 5.2 we need some estimates for $\varphi(\xi,t)$. The following result holds.

LEMMA 5.3. *Let us assume that* $(\varphi(\xi,t),R(t))$ *is a solution of* (5.4), (5.11), (5.13). *Then*

$$p(t)(R(t))^3\int_0^1\varphi(\xi,t)\xi^2d\xi=C, \tag{5.14}$$

where $C>0$ *is a suitable constant that depends on the initial data.*

Proof. Let us write

$$M=\int_0^1\varphi(\xi,t)\xi^2d\xi.$$

Multiplying (5.11) by ξ^2 and integrating in the interval $[0,1]$, we obtain

$$\begin{aligned}\tfrac{dM}{dt}&=\tfrac{A}{pR^2}\varphi_\xi(1,t)-\tfrac{A}{pR^2}\varphi_\xi(1,t)\left(\varphi(1,t)-3\int_0^1\varphi(\xi,t)\xi^2d\xi\right)\\&\quad+\left(\tfrac{1}{\gamma}-1\right)\tfrac{p_t}{p}M=\left[\tfrac{3A}{pR^2}\varphi_\xi(1,t)+\left(\tfrac{1}{\gamma}-1\right)\tfrac{p_t}{p}\right]M.\end{aligned} \tag{5.15}$$

Using (5.13) and (5.4) we deduce that

$$\begin{aligned}&\frac{3A}{pR^2}\varphi_\xi(1,t)+\left(\frac{1}{\gamma}-1\right)\frac{p_t}{p}\\&=\frac{3A}{pR^2}\left[-\frac{pR}{A}\left(1-\frac{2\sigma}{R}\frac{1}{3\gamma p}\right)R_t-\frac{pR}{A}\frac{R}{3\gamma p}p_{\infty,t}\right]+\left(\frac{1}{\gamma}-1\right)\frac{p_t}{p}\\&=-\frac{3}{R}R_t-\frac{p_t}{p},\end{aligned}$$

which combined with (5.15) implies (5.14). □

Lemma 5.3 is just a restatement of the conservation of mass of gas which can be written as

$$\int_{B_R(0)}\varrho(x,t)d^3x=M_0.$$

We now continue the proof of the the theorem. Notice that (5.4) and Lemma 5.3 imply the following. If $\varphi(\xi,t) \leq 1$ for $0 \leq \xi \leq 1$, then $R(t) \geq \delta_0 > 0$, where δ_0 depends only on $p_\infty(t)$. We now argue as follows. Notice that the function

$$\tilde{\varphi}(\xi,t) = \frac{K}{(p(t))^{1-\frac{1}{\gamma}}} \tag{5.16}$$

is a solution of (5.11). Suppose that K is selected large enough so that the following inequalities hold:

$$\tilde{\varphi}(\xi,0) > \max\{\varphi_0(\xi)\},$$

$$K > \left(\sup(p_\infty) + \frac{2\sigma}{\delta_0}\right)^\beta,$$

and

$$\tilde{\varphi}(\xi,0) > 1.$$

We can use $\tilde{\varphi}(\xi,t)$ as a supersolution for (5.11), (5.12) as far as $\tilde{\varphi}(\xi,t)$ remains larger than one. Define $t^* = \inf\{t : \tilde{\varphi}(\xi,t) = 1\}$. By assumption $t^* > 0$. Suppose that $t^* < \infty$. Then by comparison $\varphi(\xi,t^*) \leq 1$ whence $R(t^*) \geq \delta_0 > 0$, but this implies $p(t^*) \leq \sup(p_\infty) + \frac{2\sigma}{\delta_0}$. However, our assumptions on K imply $\varphi(\xi,t^*) > 1$ and this yields a contradiction with the definition of t^*.

We then obtain that $t^* = \infty$ and $\tilde{\varphi}(\xi,t)$ defined in (5.16) is a global supersolution for (5.11), (5.12). It then follows that, as far as the solution of (5.4), (5.11)–(5.13) is defined, there holds $\varphi(\xi,t) \leq \tilde{\varphi}(\xi,t)$. Using (5.14) we obtain the following estimate:

$$C \leq \frac{pR^3}{3}\tilde{\varphi}(\xi,t) = \frac{K}{3}p^{1-\beta}R^3 = \frac{K}{3}p^{\frac{1}{\gamma}}R^3. \tag{5.17}$$

Taking into account (5.4) and (5.17) as well as the fact that $\gamma > 1$, it follows that $R \geq \delta > 0$ where δ depends only on p_∞, $\varphi_0(\xi)$, $R(0)$. Using again (5.4) as well as our assumption on p_∞ we arrive at

$$0 < m < p(t) < \frac{1}{m} \tag{5.18}$$

for some suitable constant m. Taking $\varepsilon_0 > 0$ small enough we deduce that the function $\tilde{\tilde{\varphi}}(\xi,t) = \frac{\varepsilon_0}{(p(t))^\beta}$ is a subsolution for (5.11), (5.12) such that $\tilde{\tilde{\varphi}}(\xi,0) \leq \varphi_0(\xi)$. By comparison, it then follows that $\tilde{\tilde{\varphi}}(\xi,t) \leq \varphi(\xi,t)$, which combined with (5.14) yields $R(t) \leq C_1$ for some constant $C_1 > 0$. Summarizing, we have obtained the estimates

$$\Theta \leq \varphi(\xi,t) \leq \frac{1}{\Theta}, \tag{5.19}$$

$$\Theta \leq R(t) \leq \frac{1}{\Theta} \tag{5.20}$$

for some suitable constant $\Theta > 0$, as far as the functions $(\varphi(\xi,t), R(t))$ are well defined.

As a next step we obtain a lower estimate of $\varphi_\xi(1,t)$. To this end, notice that (5.4), (5.11), (5.13) imply that φ solves the problem

$$\begin{aligned}\varphi_t = &\tfrac{A}{pR^2}\Delta_\xi(\log(\varphi)) - \tfrac{A}{pR^2}\varphi_\xi(1,t)\xi\varphi_\xi \\ &- \left(1-\tfrac{1}{\gamma}\right)\left(1-\tfrac{2\sigma}{3\gamma pR}\right)^{-1}\tfrac{p_{\infty,t}}{p}\varphi + \tfrac{2\sigma 1-\frac{1}{\gamma}A}{pR^3}\left(1-\tfrac{2\sigma}{3\gamma pR}\right)^{-1}\varphi_\xi(1,t)\varphi.\end{aligned} \tag{5.21}$$

For any ε, let us define a function $u(\xi,\varepsilon)$ as

$$u(\xi,\varepsilon) = v_\varepsilon\left(\frac{\xi-1}{\varepsilon}\right), \quad y = \frac{\xi-1}{\varepsilon}, \tag{5.22}$$

where $v(y)$ solves the problem

$$(\log(v))_{yy} - v_y - \Gamma\varepsilon v = 0, \tag{5.23}$$

$$v(0) = 1, \tag{5.24}$$

$$v_y(0) = 1, \tag{5.25}$$

where $\Gamma > 0$ is a positive constant that will be fixed presently and $\varepsilon > 0$ is a small number to be precised.

It is easily seen by means of the standard theory of ODEs that

$$\lim_{\varepsilon\to 0} v_\varepsilon(y) = \frac{1}{1-y},$$

uniformly on compact sets of the y-variable. Using (5.18) and (5.20) it readily follows that taking

$$\Gamma \geq 2\sup\left\{2\sigma\left(1-\frac{1}{\gamma}\right)\left(1-\frac{2\sigma}{3\gamma pR}\right)^{-1}\right\}$$

and $\varepsilon > 0$ small enough, the following inequality holds in $-(\frac{2}{\Theta}-1)\varepsilon \leq \xi - 1 \leq 0$, where $\Theta > 0$ is as in (5.20):

$$\begin{aligned}&-\frac{A}{pR^2}\Delta_\xi(\log(u)) + \frac{A}{pR^2}u_\xi(1,t)\xi u_\xi \\ &+\frac{1}{1-\gamma}\left(1-\frac{2\sigma}{3\gamma pR}\right)^{-1}\frac{p_{\infty,t}}{p}u + \frac{2\sigma A\frac{1}{1-\gamma}}{pR^2}\left(1-\frac{2\sigma}{3\gamma pR}\right)^{-1}u_\xi(1,t)u \\ &\leq -\frac{A}{pR^2}(\log(u))_{\xi\xi} - \frac{A}{pR^2}\frac{(\log(u))_\xi}{\xi} \\ &+\frac{A}{pR^2}u_\xi(1,t)u_\xi + \frac{A}{pR^2}u_\xi(1,t)(\xi-1)u_\xi + \Gamma\frac{A}{pR^2}u_\xi(1,t)u \\ &\leq \frac{A}{pR^2}[-(\log(u))_{\xi\xi} + u_\xi(1,t)u_\xi + \Gamma u_\xi(1,t)u] = 0.\end{aligned}$$

We can use $u(\xi,\varepsilon)$ as a subsolution for (5.11), (5.12) as far as the solution of this last problem is defined. Indeed, let us select ε small enough such that $u(\xi,\varepsilon) < \varphi_0(\xi)$ in $-(\frac{2}{\Theta}-1)\varepsilon \leq \xi - 1 \leq 0$. Taking into account (5.20), it readily follows that

$u(\xi,\varepsilon) \le \varphi(\xi,t)$ at $\xi = (\frac{2}{\Theta} - 1)\varepsilon$. Notice that as far as $u(\xi,\varepsilon) \le \varphi(\xi,t)$ we have $\varphi_\xi(\xi,t) \le u_\xi(\xi,\varepsilon)$. Since $u_\xi \ge 0$, it follows that for $-(\frac{2}{\Theta} - 1)\varepsilon \le \xi - 1 \le 0$

$$\begin{aligned} &u_t - \frac{A}{pR^2}\Delta_\xi(\log(u)) + \frac{A}{pR^2}\varphi_\xi(1,t)\xi u_\xi + \left(\frac{1}{\gamma} - 1\right)\frac{p_t}{p}u \\ &\le -\frac{A}{pR^2}\Delta_\xi(\log(u)) + \frac{A}{pR^2}u_\xi(1,t)\xi u_\xi + \left(\frac{1}{\gamma} - 1\right)\left(1 - \frac{2\sigma}{3\gamma pR}\right)^{-1}\frac{p_{\infty,t}}{p}u \\ &+\frac{2\sigma A^{\frac{1}{\gamma}} - 1}{pR^2}\left(1 - \frac{2\sigma}{3\gamma pR}\right)^{-1} u_\xi(1,t)u \le 0. \end{aligned}$$

Let us define as t^* the smallest time where the inequality $\varphi_\xi(\xi,t) < u_\xi(\xi,\varepsilon)$ fails. Arguing by comparison we deduce that $u(\xi,\varepsilon) < \varphi(\xi,t^*)$, whence this inequality holds as far as $\varphi(\xi,t)$ is defined. As indicated above this implies

$$\varphi_\xi(\xi,t) \le C \tag{5.26}$$

as long as the solution of (5.4), (5.11)–(5.13) is defined. The constant $C > 0$ depends only on $\varphi_0(\xi)$, $p_\infty(t)$, $R(0)$.

To conclude the proof of Theorem 5.2, it remains only to prove a lower estimate for $\varphi_\xi(1,t)$. Such a lower estimate combined with (5.26) would imply a global existence theorem for (5.4), (5.11)–(5.13). Indeed, an estimate of $|\varphi_\xi(1,t)|$ together with (5.18)–(5.20) provides estimates for higher order derivatives of φ by means of a standard bootstrap argument applied to (5.11). Using Theorem 5.1 we can then extend the time interval of definition for (5.4), (5.11)–(5.13).

In order to obtain the desired lower estimate for $\varphi_\xi(1,t)$, we define a function $U(\xi,\varepsilon)$ as the unique solution of

$$\Delta_\xi(\log(U)) - C\xi U_\xi + \frac{\tilde{\Gamma}}{\varepsilon}CU = 0, \tag{5.27}$$

$$U(\xi,\varepsilon) = 1, \tag{5.28}$$

$$U_\xi(1,\varepsilon) = -\frac{1}{\varepsilon}, \tag{5.29}$$

where C is as in (5.26) and $\tilde{\Gamma} > 0$ will be made precise later. It is not hard to check that U behaves asymptotically as

$$U(\xi,\varepsilon) \simeq w\left(\frac{\xi - 1}{\varepsilon}\right), \quad y = \frac{\xi - 1}{\varepsilon} \tag{5.30}$$

uniformly in the region $-K\varepsilon \le \xi - 1 \le 0$, where K is any fixed positive constant, and where

$$w = e^{-y}.$$

It turns out that U is a supersolution of (5.11), (5.12) if $\tilde{\Gamma}$ is chosen large enough as far as $\varphi(\xi,t) \le U(\xi)$. Indeed, under these assumptions and also using (5.26) we have

$$-\frac{1}{\varepsilon} \le \varphi_\xi(1,t) \le C,$$

from which, if $\varepsilon > 0$ is chosen small enough,

$$U_t - \frac{A}{pR^2}\Delta_\xi(\log(U)) + \frac{A}{pR^2}\varphi_\xi(1,t)\xi U_\xi$$
$$+ \left(1 - \frac{1}{\gamma}\right)\left(1 - \frac{2\sigma}{3\gamma pR}\right)^{-1}\frac{p_{\infty,t}}{p}U + \frac{2\sigma A(1-\frac{1}{\gamma})}{pR^2}\left(1 - \frac{2\sigma}{3\gamma pR}\right)^{-1}\varphi_\xi(1,t)U$$
$$\geq -\frac{A}{pR^2}\left[\Delta_\xi(\log(U)) - C\xi U_\xi + \frac{\tilde{\Gamma}}{\varepsilon}CU\right] = 0.$$

Arguing then by comparison as in the proof of (5.26) we obtain the estimate

$$\varphi_\xi(1,t) \geq C, \tag{5.31}$$

where C depends only on $\varphi_0(\xi)$, $p_\infty(t)$, $R(0)$. Combining (5.26) and (5.31) we obtain an upper estimate of $|\varphi_\xi(1,t)| \geq C$, and this concludes the proof of Theorem 5.2. □

6. Singularity formation for $\Gamma = 0$. In this section we exhibit a singularity formation mechanism for the system (5.1)–(5.4) that can take place if the assumption $p_\infty(t) > 0$ in Theorem 5.2 is dropped. The goal here is to describe in detail one possible singularity mechanism using formal techniques. We have not attempted to provide a rigorous construction of the solutions to be exhibited. Let us assume that $p_\infty(t)$ can take negative values. For the singular solution that we describe here the function $R(t)$ remains bounded but $|\dot{R}(t)|$ blows up. An analogous phenomena occurs also for the one-dimensional undercooled Stefan problem [HV]. As in this last case the derivative $\varphi_\xi(1,t)$ blows up in finite time. However, there is a major difference between the bubble model (5.1)–(5.4) and the undercooled Stefan problem, namely, the rate of blow-up for $\varphi_\xi(1,t)$ is much larger for the Stefan problem than for the system (5.1)–(5.4). To get some intuition on the blow-up mechanism described here, we just remark that it exhibits some analogies with the blow-up mechanism that can take place for (2.19) in the case $\Gamma = 0$. In this case (2.19) reduces to an algebraic equation and the blow-up can be analyzed in a straightforward manner. After the onset of the singularity, the approximation $\Gamma = 0$ ceases to be valid. In fact, as soon as $|\dot{R}(t)|$ grows sufficiently large, the inertial terms $\Gamma(R\ddot{R} + \frac{3}{2}(\dot{R})^2)$ become relevant and should be taken into account. We have not attempted to describe the effects of those terms in the singularity formation mechanism.

For convenience we will use the equivalent formulation of (5.1)–(5.4) given by (5.4), (5.11)–(5.13). Let us denote as R_0 the value of the radius for which the singularity appears, and let T be the time of formation of the singularity. We write

$$\left(1 - \frac{2\sigma}{3\gamma pR}\right) = \frac{1}{3\gamma p}\left(3\gamma p_\infty + \frac{2\sigma(3\gamma - 1)}{R}\right). \tag{6.1}$$

The singularity that we describe here will appear due to the vanishing of the left-hand side of (6.1) (see (5.13)). Taking into account (6.1) the following relation between R_0 and T easily follows:

$$\frac{3\gamma p_\infty(T)}{3\gamma - 1} + \frac{2\sigma}{R_0} = 0. \tag{6.2}$$

Notice that at the time $t = T$ we have $p(T) = p_\infty(T) + \frac{2\sigma}{R_0} > 0$. The function p that appears in (5.11), (5.13) is then of order one near the blow-up point and it does

not affect very much the form of the singularity. By assumption, as $t \to (T)^-, R$ approaches to R_0. Thus

$$3\gamma p_\infty(t) + \frac{2\sigma(3\gamma-1)}{R} = 3\gamma(p_\infty(t) - p_\infty(T) + 2\sigma(3\gamma-1)\left(\frac{1}{R} - \frac{1}{R_0}\right)$$
$$\approx 3\gamma \dot{p}_\infty(T)(t-T) + \frac{2\sigma(3\gamma-1)}{(R_0)^2}(R_0 - R).$$

It is then natural to approximate (5.13) as

$$R_t = -\frac{\frac{3\gamma A}{R_0}\varphi_\xi(1,t) + R_0 p_{\infty,t}(T)}{3\gamma\dot{p}_\infty(T)(t-T) + \frac{2\sigma(3\gamma-1)}{(R_0)^2}(R_0 - R)}. \tag{6.3}$$

On the other hand, near the point $\xi \approx 1$ the function φ will remain close to the value one (cf. (5.12)). We then introduce the function

$$\psi = \varphi - 1. \tag{6.4}$$

Approximating $\log(\varphi)$ by ψ we can write (5.11) to the lowest order in neighborhood of $\xi = 1$ as

$$\psi_t = \frac{A}{p(T)(R_0)^2}(\Delta(\psi) - \psi_\xi(1,t)\psi_\xi) - \left(\frac{1}{\gamma} - 1\right)\frac{2\sigma R_t}{p(T)(R_0)^2}, \tag{6.5}$$

where we have approximated p_t using (5.4) by

$$p_t = p_{\infty,t} - \frac{2\sigma R_t}{(R)^2} \approx -\frac{2\sigma R_t}{(R_0)^2}.$$

Taking into account (5.12), we need to complement (6.5) with

$$\psi(1,t) = 0.$$

The blow-up mechanism that we describe here is essentially driven by the ODE (6.3) with $\varphi_\xi(1,t)$ roughly of order one (although with logarithmic corrections). It is then natural to expect a behavior of the form $(R_0 - R) \approx C\sqrt{T-t}$ or, in general, the weaker assumption

$$(R_0 - R) \gg (T-t), \tag{6.6}$$

as $t \to T$. Using (6.4) and (6.6) we can write (6.3) as

$$R_t = -\frac{R_0}{2\sigma(3\gamma-1)}\frac{3\gamma A\varphi_\xi(1,t) + (R_0)^2 p_{\infty,t}(T)}{(R_0 - R)}. \tag{6.7}$$

We now proceed to describe a singularity formation mechanism for (6.5), (6.6), (6.7) in which the resulting solution is compatible with all the approximations made. Toward this end, it is convenient to introduce self-similar variables as follows:

$$\psi(\xi,t) = (T-t)^{\frac{1}{2}}G(y,\tau), \tag{6.8}$$

$$y = \frac{\xi - 1}{\sqrt{T-t}}, \tag{6.9}$$

$$\tau = -\log(T-t), \tag{6.10}$$

$$R - R_0 = (T-t)^{\frac{1}{2}}\lambda(\tau). \tag{6.11}$$

Using this new set of variables (6.5), (6.6), (6.7) becomes

$$\begin{aligned} G_\tau &= \frac{A}{p(T)(R_0)^2}G_{yy} - \frac{y}{2}G_y \\ &+\frac{1}{2}G + \frac{2Ae^{-\tau}}{p(T)(R_0)^2(1+e^{-\tau}y)}G_y \\ &-\frac{Ae^{-\tau}}{p(T)(R_0)^2}G(0,\tau)G_y + \left(1-\frac{1}{\gamma}\right)\frac{2\sigma}{p(T)(R_0)^2}\left(\lambda_\tau - \frac{1}{2}\lambda\right), \end{aligned} \tag{6.12}$$

$$G(0,\tau) = 0, \tag{6.13}$$

$$\lambda_\tau = \frac{1}{2}\lambda + \frac{R_0}{2\sigma(3\gamma-1)}\frac{3\gamma A G_y(0,\tau) + (R_0)^2 p_{\infty,t}(T)}{\lambda}. \tag{6.14}$$

Neglecting exponentially small factors in (6.12), we obtain the approximate equation:

$$G_\tau = \frac{A}{p(T)(R_0)^2}G_{yy} - \frac{y}{2}G_y + \frac{1}{2}G + \left(1-\frac{1}{\gamma}\right)\frac{2\sigma}{p(T)(R_0)^2}\left(\lambda_\tau - \frac{1}{2}\lambda\right). \tag{6.15}$$

Problem (6.13), (6.15) does not admit stationary solutions except in the case $\lambda = 0$. In order to check this, it is enough to multiply the equation of the steady state by $y\exp(-\frac{p(T)(R_0)^2y^2}{4A})$ and integrate in the interval $(-\infty, 0)$. After integration by parts, using (6.13) it follows that $\lambda = 0$. Notice, however, that $\lambda = 0$ cannot be a steady state of (6.13)–(6.15). This excludes self-similar behaviors as possible asymptotics for singular solutions of (5.1)–(5.4). We will look then for solutions of (6.13)–(6.15) with the asymptotics

$$G(y,\tau) \sim a(\tau)y, \tag{6.16}$$

as $\tau \to \infty$. Multiplying (6.15) by $y\exp(-\frac{p(T)(R_0)^2y^2}{4A})$ and integrating in the interval $(-\infty, 0)$, we obtain after some computations the differential equation

$$a_\tau = -\frac{2\sigma(\gamma-1)}{\gamma R_0\sqrt{\pi}}\frac{(\lambda_\tau - \frac{\lambda}{2})}{\sqrt{Ap(T)}}. \tag{6.17}$$

On the other hand, (6.16) yields $G_y(0,\tau) \sim a(\tau)$ as $\tau \to \infty$. Equation (6.14) can then be approximated as

$$\lambda_\tau = \frac{\lambda}{2} + \frac{R_0}{2\sigma(3\gamma-1)}\big[3\gamma A a(\tau) + (R_0)^2 p_{\infty,t}(T)\big]\lambda. \tag{6.18}$$

It is easy to check that the system of equations (6.17), (6.18) admits solutions with the asymptotics

$$a(\tau) \sim -\frac{3\sigma(\gamma-1)^2(\tau)^2}{4\gamma(3\gamma-1)\pi R_0 p(T)} \quad \tau \to \infty, \tag{6.19}$$

$$\lambda(\tau) \sim -\frac{3(\gamma-1)\sqrt{A}\tau}{2(3\gamma-1)\sqrt{\pi R_0 p(T)}} \quad \tau \to \infty. \tag{6.20}$$

The asymptotics (6.19), (6.20) combined with (6.16) provide a solution that is compatible with all the previous approximations. Notice that in the original set of variables (6.20) implies

$$R(t) \sim R_0 - \frac{3(\gamma-1)}{2(3\gamma-1)}\sqrt{\frac{A}{\pi p(T)}}(T-t)^{\frac{1}{2}}|\log(T-t)|^{\frac{1}{2}} \quad t \to T^-. \tag{6.21}$$

Finally, we can compute the final profile of $\varphi(\xi,t)$ arguing as in [HV]. Let us fix $\bar{t}$ close enough to T and $y_0 < 0$ with $|y_0|$ large enough. According to (6.16), (6.19) there holds

$$\psi(\xi,\bar{t}) \sim -\frac{3\sigma(\gamma-1)^2|\log(T-\bar{t})|^2}{4\gamma(3\gamma-1)\pi R_0 p(T)}(\xi-1) \tag{6.22}$$

for $\xi - 1 = y_0\sqrt{T-\bar{t}}$. As in [HV], the solution of (6.5) remains approximately constant in the interval $[\bar{t},T]$ if $|y_0|$ is large enough. Taking into account that $\log(1-\xi) \sim \log(\sqrt{T-\bar{t}})$ as $\bar{t} \to T^-$, from (6.4), (6.22) the following final profile for $\varphi(\xi,t)$ then follows:

$$\varphi(\xi,t) \sim 1 + \frac{3\sigma(\gamma-1)^2(1-\xi)(\log(1-\xi))^2}{\gamma(3\gamma-1)\pi R_0 p(T)}, \quad \xi \to 1^-. \tag{6.23}$$

Formulae (6.21), (6.23) provide a rather detailed description of the singular behavior of the solution of (5.1)–(5.4). We remark that such behavior is not strictly self-similar due to the onset of logarithmic corrections.

Appendix A. In this appendix we describe some of the physical parameters that we have used in order to obtain the model (2.7), (2.8), (2.10)–(2.15). To this end, we use the experimental values in [BP2]. In the expansion phase, the radius of the bubble described in [BP2] varies between 5×10^{-4} cm. and 4×10^{-3} cm. As the average value for the radius we will take

$$R_0 \sim 2 \times 10^{-3} \text{ cm.} \tag{A.1}$$

The frequency of the external sound wave is of order $\omega = 2\pi \times 26.5$ Khz. It is then natural to introduce a characteristic time of order

$$t_{osc} \sim 4 \times 10^{-5} s. \tag{A.2}$$

The experiments in [BP2], [BWLRP] were made at approximately constant temperature. We will take a characteristic temperature T_{char} of order:

$$T_{char} \sim 300\ K. \tag{A.3}$$

The momentum equation for the gas can be written as

$$\varrho_g \frac{Dv}{Dt} = -\frac{\partial p}{\partial r} + \frac{\partial}{\partial r}\left[2\mu\frac{\partial v}{\partial r} + \left(\mu_v - \frac{2}{3}\mu\right)\nabla\cdot(\vec{v})\right] + \frac{4\mu}{r}\frac{\partial v}{\partial r} - \frac{4\mu v}{r^2}, \tag{A.4}$$

where μ, μ_v are the viscosity coefficients and ϱ_g is the density of the gas. Taking into account that $\frac{\partial p}{\partial r} = \frac{dp}{d\varrho}\frac{\partial \varrho}{\partial r}$ and that $\frac{dp}{d\varrho}$ is of the order of magnitude of the square of the velocity of the sound (that we use to denote as c), it follows that

$$\left|\frac{\partial p}{\partial r}\right| \sim \frac{c^2 \delta \varrho_g}{R_0} = \frac{c^2 \varrho_g}{R_0}\frac{\delta \varrho}{\varrho_g}, \tag{A.5}$$

where $\delta\varrho$ denotes the order of magnitude of variations of density. The order of magnitude of the speed of sound in our particular setting is

$$c^2 \sim \frac{p}{\varrho} \sim 10^9 \text{ cm}^2 s^{-2}, \tag{A.6}$$

where we have taken as an order of magnitude for the pressure the atmospheric pressure, and as an order of magnitude for the density, $\varrho \sim 10^{-3} g \text{ cm}^{-3}$. The size of the inertial terms in (A.4) is roughly

$$\varrho_g \left|\frac{Dv}{Dt}\right| \sim \varrho_g \frac{R_0}{t_{osc}^2}. \tag{A.7}$$

Finally, the order of magnitude of viscous terms in (A.4) is

$$\frac{\mu v}{R_0^2} \sim \frac{\mu}{R_0 t_{osc}}. \tag{A.8}$$

Comparison between (A.5) and (A.8) shows that viscous terms are negligible compared with pressure terms. In fact, the order of magnitude of the variations of the radius suggests that $\frac{\delta \varrho}{\varrho_g} \sim 1$. Then

$$\frac{|\frac{\partial p}{\partial r}|}{\frac{\mu v}{R_0^2}} \sim \frac{c^2 t_{osc}}{\nu},$$

where $\nu = \frac{\mu}{\varrho_g}$ stands for kinematic viscosity. At atmospheric pressure and $T = 300\, K$, ν takes the value

$$\nu = 0.160 \text{ cm}^2 s^{-1}. \tag{A.9}$$

It then follows that

$$\frac{|\frac{\partial p}{\partial r}|}{\frac{\mu v}{R_0^2}} \sim 2.5 \times 10^5. \tag{A.10}$$

On the other hand, the relative size of pressure and inertial terms has the order of magnitude

$$\frac{|\frac{\partial p}{\partial r}|}{\varrho_g t |\frac{Dv}{Dt}|} \sim \frac{c^2 t_{osc}^2}{R_0^2}\frac{\delta \varrho}{\varrho_g} \sim \frac{c^2 t_{osc}^2}{R_0^2} \sim 4 \times 10^5. \tag{A.11}$$

It follows from (A.10) and (A.11) that the leading term in (A.4) is $\frac{\partial p}{\partial r}$. As is usual in fluid mechanics, we will then approximate (A.4) by

$$\frac{\partial p}{\partial r} = 0, \tag{A.12}$$

or, in an equivalent way,

$$p = p(t). \tag{A.13}$$

In the liquid, the Reynolds number that measures the relative size of inertial and viscous terms is of order one. In fact $Re_l \sim \frac{R_0^2}{\nu_l t_{osc}} \sim 5$, where we have used that $\nu_l \sim 1\,\mathrm{cm}^2 s^{-1}$. This means that in the liquid we have to deal with the whole Navier–Stokes system. However, this is not a serious difficulty, since in the spherically symmetric case the solution of that system (including viscous terms) is given by the explicit solution (2.2).

We need to determine the relative importance of viscous terms at the boundary of the bubble. The condition of mechanical equilibrium at the surface of the bubble is

$$-p_l + p_g - \left(2\nu_l \frac{\partial v_l}{\partial r} - 2\nu_g \frac{\partial v_g}{\partial r}\right) = \frac{2\sigma}{R}. \tag{A.14}$$

It turns out that viscous terms are completely negligible compared with the effect of surface tension. In fact, using the fact that $\nu_l \sim 10^{-2} g\, cm^{-1} s^{-1} \gg \nu_g \sim 10^{-4} g\,\mathrm{cm}^{-1} s^{-1}$, it follows that

$$\nu_l \frac{\partial v_l}{\partial r} - \nu_g \frac{\partial v_g}{\partial r} \sim \frac{\nu_l}{t_{osc}}. \tag{A.15}$$

The surface tension for the air-water interface at ambient temperature is of order

$$\sigma \sim 70\, dyn\,\mathrm{cm}^{-1}. \tag{A.16}$$

Using (A.1), (A.15), (A.16), it follows that

$$\frac{|\nu_l \frac{\partial v_l}{\partial r} - \nu_g \frac{\partial v_g}{\partial r}|}{\frac{2\sigma}{R}} \sim \frac{\nu_l R}{2\sigma t_{osc}} \sim \frac{1}{300}. \tag{A.17}$$

It is then natural to assume that viscous terms are not relevant in (A.14) and we replace that equation by

$$-p_l + p_g = \frac{2\sigma}{R}. \tag{A.18}$$

We now proceed to estimate the importance of thermal effects in the model. As a first step we consider the relative size of viscous effects in the liquid. The temperature in the liquid phase satisfies the equation:

$$\rho_l c_l \frac{DT}{Dt} = \nabla(k_l \nabla T) + \nabla v : \tau', \tag{A.19}$$

where τ' is the viscous part of the stress intensity factor, and k_l, c_l are the thermal conductivity and the specific heat of the liquid, respectively.

At ambient temperature we have the following numerical values: $\alpha = \frac{k_l}{\varrho_l c_l} \sim 1.5 \times 10^{-3}\ \mathrm{cm}^2 s^{-1}$, $c_l \sim 4 \times 10^7\, erg\, g^{-1} K^{-1}$. We then obtain the following relative sizes for the different terms in (A.19):

$$\frac{|\nabla(k_l \nabla T)|}{\rho_l c_l |\frac{DT}{Dt}|} \sim \alpha \frac{t_{osc}}{R_0^2} = 1.5 \times 10^{-2}, \tag{A.20}$$

$$\frac{|\nabla v : \tau'|}{\rho_l c_l |\frac{DT}{Dt}|} \sim \frac{\nu_l}{\rho_l c_l t_{osc} T_l} \left(\frac{\delta T}{T_l}\right)^{-1} \sim \frac{1}{4.8 \times 10^5} \left(\frac{\delta T}{T_l}\right)^{-1}. \tag{A.21}$$

Notice that (A.21) indicates that viscous terms are negligible as soon as relative changes of temperature are larger that 10^{-5}. It follows from (A.20), (A.21) that we can approximate (A.19) as

$$\frac{DT}{Dt} = 0. \tag{A.22}$$

Assuming that the initial temperature of the liquid is constant ($T = T_l$), we then deduce that T remains constant everywhere. Using the fact that the temperature is continuous across the interface, we then obtain the following boundary condition for the temperature of the gas at the surface of the bubble:

$$T = T_l \ \text{ at } \ r = R(t). \tag{A.23}$$

As a next step we proceed to determine the relevant terms in the thermal equation for the gas. The entropy equation for the gas can be written as

$$\varrho T \frac{Ds}{Dt} = \nabla(k\nabla T) + \nabla v : \tau'. \tag{A.24}$$

We can readily compare the sizes of the viscous and conduction terms:

$$\frac{|\nabla v : \tau'|}{\nabla(k\nabla T)} \sim \frac{v_g R_0^2}{\alpha c_v T} \sim 10^{-15},$$

where we have made the assumption $\frac{\delta T}{T} \sim 1$. The effect of viscous terms in (A.24) is then completely negligible. On the other hand, we need to compare convective and conductive terms in (A.24). The relative size of these terms is given by

$$\frac{|\varrho T \frac{Ds}{Dt}|}{|\nabla(k\nabla T)|} \sim \frac{R_0^2}{\alpha t_{osc}} \sim \frac{1}{2}. \tag{A.25}$$

We are thus led to assume that convective and conductive terms in (A.24) are of the same order of magnitude. We then replace (A.24) by

$$\rho T \frac{Ds}{Dt} = \nabla(k\nabla T). \tag{A.26}$$

Equation (A.26) has to be complemented with the boundary condition (A.23).

To conclude this appendix we adimensionalize the resulting model using R_0 as the unit of length, t_{osc} as the unit of time, using as characteristic units of density and pressure on the liquid, respectively, $\rho_s = 10^{-3} g\ \text{cm}^{-3}, p_0 = 1$ atm. Finally, we use as the unit of density in the liquid $\varrho_l = 1\, g\ \text{cm}^{-3}$. With the above mentioned approximations, the evolution of the bubble turns out to be described for the model (2.19)–(2.22), where we have to take into account that in (2.19) σ stands for $\frac{\sigma}{R_0 p_0}$ and in (2.22) k is the nondimensional number $\frac{\alpha t_{osc}}{R_0^2}$. It is relevant to determine the order of magnitude of the number Γ in (2.19), which is given by

$$\Gamma = \frac{R_0^2 \varrho_l}{p_0 t_c^2} \sim \frac{1}{400}.$$

Appendix B. We recall here the derivation of (4.4) in the radially symmetric case. Notice that

$$\frac{d}{dt}\left(\int_{B_{R(t)}} \varrho\Phi d^3x\right) = \frac{d}{dt}\int_0^{R(t)} 4\pi\varrho\Phi r^2 dr$$
$$= 4\pi R^2\varrho\Phi|_{R(t)}\dot{R} + \int_0^{R(t)} 4\pi(\varrho_t\Phi + \varrho\Phi_t)r^2 dr.$$

Taking into account the continuity equation (2.18) as well as the boundary condition (2.15) we obtain

$$\frac{d}{dt}\left(\int_{B_{R(t)}} \varrho\Phi d^3x\right) = 4\pi R^2(\varrho\Phi v)|_{R(t)} + 4\pi\int_0^{R(t)} \varrho\Phi_t r^2 dr$$
$$-4\pi\int_0^{R(t)} \frac{\partial}{\partial r}(r^2\varrho v)\Phi dr = 4\pi\int_0^{R(t)} \varrho\frac{D\Phi}{Dt}r^2 dr = \int_{B_{R(t)}} \varrho\frac{D\Phi}{Dt}d^3x,$$

as we wanted to prove.

Acknowledgments. The authors are very thankful to A. Liñán for several interesting discussions about the formulation of the model considered in this paper. Part of this work was made during a stay by the first author at Universidad Complutense of Madrid funded by the European Science Foundation through his FBP Scientific Programme.

REFERENCES

[BP1] B.P. Barber and S.J. Putterman, *Observation of synchronous picosecond sonoluminescence*, Letters to Nature, 9 (1991), pp. 318–323.

[BP2] B.P. Barber and S.J. Putterman, *Light scattering measurements of the repetitive supersonic implosion of a sonoluminescing bubble*, Phys. Rev. Lett., 26 (1992), pp. 3839–3841.

[BWLRP] B.P. Barber, C.C. Wu, R. Löfstedt, P.H. Roberts, and S.J. Putterman, *Sensitivity of sonoluminescence to experimental parameters*, Phys. Rev. Lett., 9 (1994), pp. 1380–1383.

[Bt] G.K. Batchelor, *An Introduction to Fluid Dynamics*, Cambridge University Press, Cambridge, UK, 1994.

[Fr] A. Friedman, *Partial Differential Equations of Parabolic Type*, Prentice Hall, Englewood Cliffs, NJ, 1964.

[HV] M.A. Herrero and J.J.L. Velázquez, *Singularity formation in the one–dimensional supercooled Stefan problem*, European J. Appl. Math., 7 (1996), pp. 119–150.

[LL] L.D. Landau and E.M. Lifshitz, *Statistical Physics*, Vol. 5, 3rd ed., Pergamon, Oxford, 1994, pp. 10–20.

[Le] T.G. Leighton, *The Acoustic Bubble*, Academic Press, New York, 1994.

[Ra] Lord Rayleigh, *On the pressure developed in a liquid during the collapse of a spherical cavity*, Phil. Mag., (1917), pp. 94–98.

SIAM J. MATH. ANAL.
Vol. 32, No. 1, pp. 172–193

A SOLUTION WITH BOUNDED EXPANSION RATE TO THE MODEL OF VISCOUS PRESSURELESS GASES*

LAURENT BOUDIN†

Abstract. We prove the existence of a global smooth solution to the equations of one-dimensional viscous pressureless gases and obtain a uniform upper estimate on the expansion rate with respect to viscosity. Then we get the convergence to the inviscid model in the duality sense of Bouchut and James.

Key words. pressureless gases, viscous solutions, entropy conditions, Oleinik entropy condition, inviscid limit, duality solutions

AMS subject classifications. 35L70, 76D05, 35R05

PII. S0036141098346840

1. Introduction and main results. We study here the one-dimensional system describing a viscous pressureless gas. The gas density $\rho(t,x) > 0$ and the velocity $u(t,x) \in \mathbb{R}$ have to satisfy the following equations:

$$\partial_t \rho + \partial_x(\rho u) = 0 \qquad \text{in }]0,T[\times\mathbb{R}, \tag{1.1}$$

$$\partial_t u + u\partial_x u = \varepsilon \frac{\partial^2_{xx} u}{\rho} \qquad \text{in }]0,T[\times\mathbb{R}, \tag{1.2}$$

where $T > 0$, and $\varepsilon > 0$ is the viscosity coefficient. We also give initial conditions

$$u(0,.) = u_0, \qquad \rho(0,.) = \rho_0. \tag{1.3}$$

We could also study the same problem set in $]0,+\infty[\times\mathbb{R}$. All the results in this paper can be extended to that case.

Remark. Note that (1.1)–(1.2) formally imply

$$\partial_t(\rho u) + \partial_x(\rho u^2) = \varepsilon \partial^2_{xx} u \qquad \text{in }]0,T[\times\mathbb{R}. \tag{1.4}$$

This computation is justified by the smoothness of ρ and u. When $\varepsilon = 0$, (1.4) is the momentum conservation law, i.e., we find the model of inviscid pressureless gases.

The equations can be seen as a simplified model of Navier–Stokes equations of gas dynamics, where the pressure has been set to 0. They can describe either cold plasmas or galaxies' dynamics [16]. One-dimensional Navier–Stokes equations with pressure are studied, for example, in [13] or [11] and the references therein.

Several results have been obtained recently on the inviscid pressureless problem by Grenier [9], Weinan, Rykov, and Sinai [7], Brenier and Grenier [5], Bouchut [1], and Bouchut and James [2], [3], [4] (see also Poupaud and Rascle [14] for an approach in the multidimensional case). The main feature that comes out is the importance of the estimate $\partial_x u \leq 1/t$ on the expansion rate. We prove here a similar estimate on the viscous problem.

*Received by the editors November 6, 1998; accepted for publication September 23, 1999; published electronically June 22, 2000.

http://www.siam.org/journals/sima/32-1/34684.html

†Laboratoire de Mathématiques, Applications, et Physique Mathématiques d' Orléans, Université d' Orléans, UMR CNRS 6628, BP 6759, 45067 Orléans Cedex 2, France (boudin@labomath.univ-orleans.fr).

We assume that initial data $\rho_0 > 0$ and u_0 satisfy the following conditions:

$$\rho_0 \in L^\infty(\mathbb{R}), \frac{1}{\rho_0} \in L^\infty(\mathbb{R}), \qquad \partial_x \rho_0 \in L^1 \cap L^\infty(\mathbb{R}), \tag{1.5}$$

$$u_0 \in L^1 \cap L^2(\mathbb{R}), \qquad \partial_x u_0 \in L^1 \cap L^2(\mathbb{R}). \tag{1.6}$$

Notations. We denote by $C_0(\mathbb{R})$ the space of continuous functions of $x \in \mathbb{R}$ that tend to 0 when $|x| \to +\infty$. Then $\mathcal{D}_+(I)$ denotes the space of nonnegative functions in $\mathcal{D}(I)$, for any open interval $I \subset \mathbb{R}$. Finally, E is the Banach space

$$E = \{v \in L^2_t(C_{0_x}) \ \mid \ \partial_x v \in L^2_t(C_{0_x}) \cap L^2_t(L^1_x)\}, \tag{1.7}$$

with the norm

$$\|v\|_E = \|v\|_{L^2_t(C_{0_x})} + \|\partial_x v\|_{L^2_t(C_{0_x})} + \|\partial_x v\|_{L^2_t(L^1_x)}.$$

Our first main result follows.

THEOREM 1. *Let T, $\varepsilon > 0$ and two functions $\rho_0 > 0$ and u_0 satisfying* (1.5) *and* (1.6). *Then there exists a solution (ρ, u) in the sense of distributions to* (1.1)–(1.3) *such that $\rho > 0$,*

$$\rho \in C([0,T]; L^1_{\mathrm{loc}}(\mathbb{R})) \cap L^\infty(]0,T[\times\mathbb{R}), \qquad \frac{1}{\rho} \in L^\infty(]0,T[\times\mathbb{R}), \tag{1.8}$$

$$u \in C([0,T]; H^1(\mathbb{R})) \cap L^2(]0,T[; H^2(\mathbb{R})). \tag{1.9}$$

Moreover, the following a priori estimates hold, with $w = \partial_x u$, $w_0 = \partial_x u_0$:

$$\|w(t)\|_{L^1_x} \le \|w_0\|_{L^1}, \quad 0 \le t \le T, \qquad \text{and} \qquad \|u\|_{L^\infty_{t,x}} \le \|u_0\|_{L^\infty}.$$

We also have, in $\mathcal{D}'_{t,x}$, the following renormalized equations for any $S \in C^2(\mathbb{R})$ such that $S'' \in L^\infty$:

$$\partial_t(\rho S(u)) + \partial_x(\rho u S(u)) - \varepsilon \partial_x(S'(u)\partial_x u) = -\varepsilon S''(u)(\partial_x u)^2, \tag{1.10}$$

$$\begin{aligned} &\partial_t S(w) + \partial_x(uS(w)) - \varepsilon \partial_x\left(\frac{S'(w)\partial_x w}{\rho}\right) \\ &\qquad = w(S(w) - wS'(w)) - \varepsilon S''(w)\frac{(\partial_x w)^2}{\rho}. \end{aligned} \tag{1.11}$$

Finally, if we assume that $\operatorname{ess\,sup} \partial_x u_0 < +\infty$, *we have the uniform upper bound estimate on the expansion rate*

$$\partial_x u(t,x) \le \frac{A}{At+1} \le \frac{1}{t} \qquad \text{almost everywhere (a.e.) } (t,x) \in [0,T]\times\mathbb{R}, \tag{1.12}$$

where $A = \max(\operatorname{ess\,sup} \partial_x u_0, 0)$.

Remark. For a given u, the solution ρ to (1.1) can be obtained with the results of DiPerna and Lions [6] or Bouchut and James [2]. Note that the norms $\|\rho\|_{L^\infty_{t,x}}$ and $\|1/\rho\|_{L^\infty_{t,x}}$ can be estimated in terms of $\|\rho_0\|_{L^\infty}$, $\|1/\rho_0\|_{L^\infty}$, and $\|u\|_E$ (see formula (4.1)).

The renormalized equations (1.10)–(1.11) give entropy inequalities when S is convex, and the inequality (1.12) on $\partial_x u$ is called the Oleinik entropy condition. It has

been proved in its optimal form by Hoff [10] for the solution of a scalar conservation law.

The proof of Theorem 1 is divided into three steps. We first solve the linear equation associated to (1.2), namely,

$$\partial_t u + b\partial_x u = \sigma \partial^2_{xx} u, \tag{1.13}$$

where b and σ are fixed, by using a standard theorem of J.-L. Lions [12], [15]. Next we use Schäfer's fixed point theorem [8], applied as follows. We start with a smooth enough approximate velocity $\tilde{u}$ and get the solution $\rho > 0$ to

$$\partial_t \rho + \partial_x(\rho \tilde{u}) = 0.$$

Then we solve (1.13) with $b = \tilde{u}$ and $\sigma = \varepsilon/\rho$. Thus we build u from $\tilde{u}$. We conclude with Schäfer's result applied to the operator $Q : \tilde{u} \mapsto u$. Then we get estimate (1.12).

Finally, we prove that condition (1.12) is sufficient to justify the asymptotics $\varepsilon \to 0$, in the sense defined by Bouchut and James [2], [3], [4].

Definition 1. *A solution $p \in \mathrm{Lip}_{\mathrm{loc}}([0,T] \times \mathbb{R})$ to*

$$\partial_t p + u\partial_x p = 0 \qquad \text{in }]0,T[\times\mathbb{R} \tag{1.14}$$

is said to be reversible if there exists two solutions $p_1, p_2 \in \mathrm{Lip}_{\mathrm{loc}_{t,x}}$ to (1.14) such that

$$\partial_x p_1 \geq 0, \qquad \partial_x p_2 \geq 0, \qquad \text{and} \qquad p = p_1 - p_2.$$

Remark. The backward problem (1.14) with final Cauchy data $p_T \in \mathrm{Lip}_{\mathrm{loc}}(\mathbb{R})$ is well-posed in the class of reversible solutions if we assume that $u \in L^\infty_{t,x}$ satisfies the Oleinik entropy condition $\partial_x u \leq 1/t$.

Definition 2. *We say that $\mu \in C([0,T]; w^*\text{-}\mathcal{M}_{\mathrm{loc}_x})$ is a duality solution to*

$$\partial_t \mu + \partial_x(\mu u) = 0 \qquad \text{in }]0,T[\times\mathbb{R} \tag{1.15}$$

if, for any $0 < \tau \leq T$, and any reversible solution p, with compact support in x, to $\partial_t p + u\partial_x p = 0$ in $]0,\tau[\times\mathbb{R}$, the function

$$t \mapsto \int_{\mathbb{R}} p(t,x)\mu(t,dx)$$

is constant on $[0,\tau]$.

For our viscous system

$$\partial_t \rho^\varepsilon + \partial_x(\rho^\varepsilon u^\varepsilon) = 0, \tag{1.16}$$

$$\partial_t q^\varepsilon + \partial_x(q^\varepsilon u^\varepsilon) = \varepsilon \partial^2_{xx} u^\varepsilon, \tag{1.17}$$

with $q^\varepsilon = \rho^\varepsilon u^\varepsilon$, the notion of duality solution enables us to justify the convergence when $\varepsilon \to 0$ to

$$\partial_t \rho + \partial_x(\rho u) = 0, \tag{1.18}$$

$$\partial_t q + \partial_x(q u) = 0, \tag{1.19}$$

with $q = \rho u$, i.e., the system of inviscid pressureless gases. We have the following result.

THEOREM 2. *We assume that* ρ_0^ε *and* u_0^ε *satisfy* (1.5)–(1.6) *for each* ε *and*

$$(q_0^\varepsilon) \rightharpoonup q_0 \text{ in } w^*\text{-}\mathcal{M}_{\mathrm{loc}}, \tag{1.20}$$
$$\rho_0^\varepsilon > 0, \qquad \|1/\rho_0^\varepsilon\|_{L^\infty} \le \frac{C}{\varepsilon^{1/4}}, \qquad (\rho_0^\varepsilon) \rightharpoonup \rho_0 \ge 0 \text{ in } \mathrm{w}^*\text{-}\mathcal{M}_{\mathrm{loc}},$$
$$\|u_0^\varepsilon\|_{L^\infty} \le C, \qquad \text{ess sup}\, \partial_x u_0^\varepsilon \le \frac{C}{\sqrt{\varepsilon}},$$

where C *is a constant independent on* ε. *Then there exists a subsequence of* $(\rho^\varepsilon, q^\varepsilon)$ *solutions to* (1.16)–(1.17) *with initial data* $(\rho_0^\varepsilon, q_0^\varepsilon)$ *which converge in* $C_t(w^*\text{-}\mathcal{M}_{\mathrm{loc}_x})$ *to* (ρ,q) *solution to* (1.18)–(1.19) *in the sense of duality, in any subinterval* $]t_1, t_2[$, $0 < t_1 < t_2 \le T$, *with initial data* ρ_0 *and* q_0, *and where* u *is a limit of a subsequence of* (u^ε) *in* $w^*\text{-}L^\infty$.

This result is obtained via a backward approximate viscous problem, which solutions, up to a subsequence, converge towards the reversible solutions to the inviscid problem.

Remark. For simplicity, in this theorem and its proof, we identify u and its universal representative defined in [3].

2. Linear equation—a priori estimates. Let three functions b, σ, and u_0 be such that

$$b \in E \qquad \text{and} \qquad \|b\|_{L^\infty_{t,x}} \le b_M, \tag{2.1}$$
$$\sigma_m \le \sigma(t,x) \le \sigma_M \qquad \text{a.e. } (t,x) \in [0,T] \times \mathbb{R}, \tag{2.2}$$
$$u_0 \in L^2(\mathbb{R}), \qquad \partial_x u_0 \in L^1 \cap L^2(\mathbb{R}), \tag{2.3}$$

where b_M, σ_m, and σ_M are strictly positive constants.

In this section, we prove the existence of a solution u to

$$\partial_t u + b\partial_x u = \sigma \partial^2_{xx} u \qquad \text{in }]0,T[\times\mathbb{R}, \tag{2.4}$$
$$u(0,.) = u_0. \tag{2.5}$$

It is obtained via the equation on $\partial_x u$.

PROPOSITION 3. *If* b, σ *satisfy* (2.1)–(2.2) *and for any*

$$w_0 \in L^1 \cap L^2(\mathbb{R}), \tag{2.6}$$

there exists a unique solution in the sense of distributions

$$w \in C_t(L^2_x) \cap L^2_t(H^1_x) \tag{2.7}$$

to

$$\partial_t w + \partial_x(bw) = \partial_x(\sigma \partial_x w) \qquad \text{in }]0,T[\times\mathbb{R}, \tag{2.8}$$
$$w(0,.) = w_0. \tag{2.9}$$

The problem (2.8)–(2.9) *satisfies the following property of stability, denoted by* (S). *Let us consider* (b_n), (σ_n), *and* (w_{0_n}) *satisfying* (2.1), (2.2), (2.6) *for each* n *and such that*

$$(b_n) \to b \text{ in } E, \tag{2.10}$$
$$(\sigma_n(t,x)) \to \sigma(t,x) \text{ a.e.}, \tag{2.11}$$
$$(w_{0_n}) \to w_0 \text{ in } L^2(\mathbb{R}). \tag{2.12}$$

Then the unique sequence (w_n) *defined by*

$$\partial_t w_n + \partial_x(b_n w_n) = \partial_x(\sigma_n \partial_x w_n) \qquad \text{in }]0,T[\times\mathbb{R}, \tag{2.13}$$

$$w_n(0,.) = w_{0_n} \tag{2.14}$$

strongly converges to the solution w *to* (2.8)–(2.9) *in* $C_t(L^2_x) \cap L^2_t(H^1_x)$.

We also have the renormalized equation on w *in* $\mathcal{D}'_{t,x}$

$$\begin{aligned} &\partial_t(S(w)) + \partial_x(bS(w)) - \partial_x(\sigma S'(w)\partial_x w) \\ &\qquad = \partial_x b(S(w) - wS'(w)) - \sigma S''(w)(\partial_x w)^2 \end{aligned} \tag{2.15}$$

for any $S \in C^2(\mathbb{R})$ *such that* S'' *is bounded.*

Finally we have

$$w \in C_t(L^1_x) \qquad \text{and} \qquad \|w(t)\|_{L^1_x} \le \|w_0\|_{L^1}, \qquad 0 \le t \le T, \tag{2.16}$$

and the estimate

$$\|w\|_{L^2_t(C_{0_x})} \le C\|\partial_x w\|^{2/3}_{L^2_{t,x}} \tag{2.17}$$

holds with a constant C *only depending on* $\|w_0\|_{L^1}$.

Once Proposition 3 is proved, we can easily obtain the following proposition.

PROPOSITION 4. *If* b, σ, *and* u_0 *satisfy* (2.1)–(2.3), *it is possible to define* $u \in C_t(C_{0_x})$ *by*

$$u(t,x) = \int_{-\infty}^{x} w(t,y)dy, \tag{2.18}$$

where w *is the solution to* (2.8) *given by Proposition* 3 *with initial data* $\partial_x u_0$. *Then* u *satisfies* (2.4)–(2.5) *and the following maximum principle estimate:*

$$\|u\|_{C_t(C_{0_x})} \le \|w_0\|_{L^1}. \tag{2.19}$$

In order to prove existence and uniqueness of a solution w to (2.8), we recall the following.

THEOREM 5 (J.-L. Lions). *Let* $(H, |.|_H)$ *and* $(V, |.|_V)$ *be two Hilbert spaces such that*

$$V \subset H \subset V'$$

with continuous and dense injections.

For a.e. $t \in [0,T]$, *we consider a bilinear form* $a(t;.,.) : V^2 \to \mathbb{R}$ *satisfying*

(i) $t \mapsto a(t;w,v)$ *is measurable for any* $(w,v) \in V^2$,

(ii) $|a(t;w,v)| \le M|w|_V|v|_V$ *a.e.* $t \in [0,T]$ $\forall w,v \in V$,

(iii) $a(t;v,v) \ge \alpha|v|^2_V - C|v|^2_H$ *a.e.* $t \in [0,T]$ $\forall w,v \in V$,

where $M, C \in \mathbb{R}$ *and* $\alpha > 0$ *are constants.*

Then, for $f \in L^2(]0,T[;V')$, $w_0 \in H$, *there exists a unique* w *such that*

$$w \in C([0,T];H) \cap L^2(]0,T[;V), \qquad \partial_t w \in L^2(]0,T[;V'),$$

and

$$\left\{ \begin{array}{rcll} \langle \partial_t w, v\rangle + a(t;w,v) & = & \langle f(t), v\rangle & \text{a.e. } t \quad \forall v \in V, \\ w(0,.) & = & w_0. & \end{array} \right.$$

This result can be found in [12] or [15].

Proof of Proposition 3. We apply Theorem 5 with $H = L^2_x(\mathbb{R})$ and $V = H^1_x(\mathbb{R})$ with their usual norm, and set

$$a(t; w, v) = -\int_{\mathbb{R}} bw\partial_x v dx + \int_{\mathbb{R}} \sigma\partial_x w\partial_x v dx \qquad \text{a.e. } t \quad \forall w, v \in H^1_x.$$

Then $a(t; ., .)$ is clearly a bilinear form on H^1_x and $t \mapsto a(t; w, v)$ is measurable for any w, v, because b and σ are measurable too. Condition (ii) is clear. For condition (iii),

$$a(t; v, v) \geq \sigma_m \|v\|^2_{H^1_x} - \sigma_m \|v\|^2_{L^2_x} - b_M \|v\|_{L^2_x} \|v\|_{H^1_x}.$$

Let us choose a constant α such that $\alpha > b_M/2\sigma_m$. Then the constant $\sigma_m - b_M/2\alpha$ is strictly positive, and since, for any v,

$$a(t; v, v) \geq \left(\sigma_m - \frac{b_M}{2\alpha}\right)\|v\|^2_{H^1_x} - \left(\sigma_m + \frac{b_M\alpha}{2}\right)\|v\|^2_{L^2_x},$$

we get (iii).

Therefore Theorem 5 gives the existence and uniqueness of the solution w to (2.7)–(2.9) and

$$\langle \partial_t w, v\rangle + a(t; w, v) = 0 \qquad \text{a.e. } t \quad \forall v \in V. \tag{2.20}$$

Note that the solution to (2.20) is also a solution to (2.8) in the sense of distributions. The inverse holds too, thanks to the density of tensor products of functions of $\mathcal{D}(]0, T[)$ and $\mathcal{D}(\mathbb{R})$ in $\mathcal{D}(]0, T[\times\mathbb{R})$. This ends the proof of existence and uniqueness for Proposition 3. □

Proof of (S). Let us prove first some preliminary results.

LEMMA 1. *For any* $(t_1, t_2) \in [0, T]^2$, *we have*

$$\frac{1}{2}\left[\int_{\mathbb{R}} (w(t_2)^2 - w(t_1)^2)dx\right] = \int_{t_1}^{t_2}\int_{\mathbb{R}} (bw\partial_x w - \sigma(\partial_x w)^2)dxdt. \tag{2.21}$$

Proof. The estimate is easily obtained by considering the equation satisfied by $\theta_m * w$, where $(\theta_m(x))$ is a mollifying sequence. □

LEMMA 2. *If we set*

$$A_1 = \|w_0\|_{L^2} \ \exp\left(\frac{1}{2}\|\partial_x b\|_{L^2_t(C_{0_x})}\sqrt{T}\right), \tag{2.22}$$

$$A_2 = \left[\frac{1}{2\sigma_m}\left(\|w_0\|^2_{L^2} + A_1^2\|\partial_x b\|_{L^2_t(C_{0_x})}\sqrt{T}\right)\right]^{1/2}, \tag{2.23}$$

$$A_3 = \sqrt{2}((b_M A_1)^2 T + (\sigma_M A_2)^2)^{1/2}, \tag{2.24}$$

then we have

$$\|w(t)\|_{L^2_x} \leq A_1, \qquad 0 \leq t \leq T, \tag{2.25}$$

$$\|\partial_x w\|_{L^2_{t,x}} \leq A_2, \tag{2.26}$$

$$\|\partial_t w\|_{L^2_t(H^{-1}_x)} \leq A_3. \tag{2.27}$$

Proof. For (2.25), we use (2.21) with $t_1 = 0$ and $t_2 = t \in]0, T]$; that is,

$$\int_{\mathbb{R}} \frac{w(t)^2}{2} dx - \int_{\mathbb{R}} \frac{w_0^2}{2} dx = \int_0^t \int_{\mathbb{R}} \left[bw\partial_x w - \sigma(\partial_x w)^2 \right] dx ds. \tag{2.28}$$

This implies

$$\|w(t)\|_{L^2_x}^2 \leq \|w_0\|_{L^2_x}^2 + \int_0^t \|w(s)\|_{L^2_x}^2 \|\partial_x b(s)\|_{C_{0_x}} ds.$$

Thanks to Gronwall's lemma, we get (2.25).

For (2.26), we use (2.21) with $t_1 = 0$ and $t_2 = T$; that is,

$$\int_{\mathbb{R}} \frac{w(T)^2}{2} dx - \int_{\mathbb{R}} \frac{w_0^2}{2} dx = - \int_0^T \int_{\mathbb{R}} \left[\partial_x b \frac{w^2}{2} + \sigma(\partial_x w)^2 \right] dx dt.$$

Then

$$\int_0^T \int_{\mathbb{R}} \sigma(\partial_x w)^2 dx dt \leq \frac{1}{2} \left[\|w_0\|_{L^2}^2 - \int_0^T \int_{\mathbb{R}} \partial_x b \; w^2 dx dt \right] \tag{2.29}$$

$$\leq \frac{1}{2} \left(\|w_0\|_{L^2}^2 + A_1^2 \|\partial_x b\|_{L^2_t(C_{0_x})} \sqrt{T} \right), \tag{2.30}$$

and we get (2.26).

Let us prove (2.27). For any $\varphi \in H^1_x$ and a.e. t,

$$\langle \partial_t w(t), \varphi \rangle_{H_x^{-1}, H_x^1} = \langle b(t) w(t) - \sigma(t) \partial_x w(t), \varphi' \rangle_{L^2_x, L^2_x}.$$

Hence, by (2.25), (2.26), and after integration, we get (2.27). □

We now begin the proof of (S). We want to prove that the sequence given by the solutions w_n to (2.13)–(2.14) strongly converges to the solution w to (2.8)–(2.9) in $C_t(L^2_x) \cap L^2_t(H^1_x)$.

First, we prove that

$$(w_n) \to w \text{ in } C_t(\text{w-}L^2_x), \tag{2.31}$$

and

$$(\partial_x w_n) \rightharpoonup \partial_x w \text{ in w} - L^2_{t,x}. \tag{2.32}$$

Thanks to Lemma 2, we know for each n and any t that

$$\|w_n(t)\|_{L^2_x} \leq \|w_{0_n}\|_{L^2} \; \exp\left(\frac{1}{2} \|\partial_x b_n\|_{L^2_t(C_{0_x})} \sqrt{T} \right).$$

From (2.10) and (2.12), we can state that there exists a constant B_1 depending only on b and w_0 such that

$$\|w_n(t)\|_{L^2_x} \leq B_1, \qquad 0 \leq t \leq T, \qquad \forall n. \tag{2.33}$$

In the same way, we can find constants B_2 and B_3 such that

$$\|\partial_x w_n\|_{L^2_{t,x}} \leq B_2 \qquad \forall n, \tag{2.34}$$

$$\|\partial_t w_n\|_{L^2_t(H^{-1}_x)} \leq B_3 \qquad \forall n. \tag{2.35}$$

Thanks to Ascoli's theorem applied in the closed ball $\mathcal{K}$, which is a metric compact space for the weak topology of $L^2(\mathbb{R})$, we can state that there exists $\omega \in C_t(\text{w-}L^2_x)$ such that, up to a subsequence,

$$(w_n) \to \omega \text{ in } C_t(\text{w-}L^2_x).$$

Furthermore, by (2.34), we can state that $\omega \in L^2_t(H^1_x)$ and, up to a subsequence,

$$(\partial_x w_n) \rightharpoonup \partial_x \omega \text{ in } \text{w} - L^2_{t,x}.$$

In (2.13) and (2.14), it is then possible to let $n \to \infty$ and we find that (2.8) and (2.9) are verified by ω. We now must prove that $\omega = w$.

LEMMA 3. *If $z \in C_t(\text{w} - L^2_x) \cap L^2_t(H^1_x)$ is a solution in the sense of distributions to (2.8)–(2.9), then $z \in C_t(L^2_x)$.*

Proof. We use the same argument as in the proof of (2.21) with the mollifying sequence (θ_m) and $z_m = \theta_m *_x z$. We then prove that $t \mapsto \|z(t)\|^2_{L^2_x}$ is continuous. Since $z \in C_t(\text{w} - L^2_x)$, the proof is ended. □

Lemma 3 ensures that $\omega \in C_t(L^2_x)$. Thanks to the uniqueness of a solution to (2.8)–(2.9) in $C_t(L^2_x) \cap L^2_t(H^1_x)$ given in Theorem 5, we find that $\omega = w$ and (2.31) and (2.32) are proved.

We note here that (2.32), (2.34), and (2.35) imply that, up to a subsequence,

$$(w_n(t,x)) \to w(t,x) \qquad \text{a.e.} \tag{2.36}$$

Let us now prove that the weak convergences (2.31) and (2.32) are strong. By (2.21), we have, for any $0 \leq t \leq T$,

$$\begin{aligned}\int_{\mathbb{R}} w_n(t)^2 dx - \int_{\mathbb{R}} w(t)^2 dx &= \int_{\mathbb{R}} w^2_{0_n} dx - \int_{\mathbb{R}} w_0^2 dx \\ &+ \int_0^t \int_{\mathbb{R}} (\partial_x b w^2 - \partial_x b_n w_n^2) dx ds + 2\int_0^t \int_{\mathbb{R}} \left(\sigma(\partial_x w)^2 - \sigma_n(\partial_x w_n)^2\right) dx ds.\end{aligned} \tag{2.37}$$

Let us study the term

$$\int_0^t \int_{\mathbb{R}} (\partial_x b w^2 - \partial_x b_n w_n^2) dx ds.$$

By interpolation, with (2.33) and (2.34), we find a constant B_5 such that

$$\|w_n\|_{L^2_t(C_{0_x})} \leq B_5 \qquad \forall n, \tag{2.38}$$

and once again by interpolation, with (2.33) and (2.38), we find a constant $B_{4,p}$ such that, for any $p \in [2, \infty]$,

$$\|w_n\|_{L^{2p/p-2}_t(L^p_x)} \leq B_{4,p} \qquad \forall n. \tag{2.39}$$

The conjunction of (2.36) and (2.39) implies that, up to a subsequence,

$$(w_n^2) \to w^2 \text{ in } L^2_t(L^{p/2}_x) \tag{2.40}$$

for $2 < p < 4$.

Besides, since $(\partial_x b_n) \to \partial_x b$ in both $L^2_t(C_{0_x})$ and $L^2_t(L^1_x)$, we get

$$(\partial_x b_n) \to \partial_x b \text{ in } L^2_t(L^{p/p-2}_x). \tag{2.41}$$

From (2.40) and (2.41), we can state that

$$(\partial_x b_n w_n^2) \to \partial_x b w^2 \text{ in } L^1_{t,x}. \tag{2.42}$$

Moreover, by (2.12), $(\|w_{0_n}\|_{L^2}) \to \|w_0\|_{L^2}$, so (2.37) can be transformed, for any t, into

$$\begin{aligned} &\frac{1}{2}\left(\|w_n(t)\|^2_{L^2_x} - \|w(t)\|^2_{L^2_x}\right) \\ &\qquad = \varepsilon_n(t) + \int_0^t \int_{\mathbb{R}} \left(\sigma(\partial_x w)^2 - \sigma_n(\partial_x w_n)^2\right) dxds, \end{aligned} \tag{2.43}$$

where (ε_n) is a sequence of functions of t that uniformly tends to 0, because of (2.42).

Let $t \in [0,T]$ be fixed. We know (2.31), and by (2.29), we can state that, up to a subsequence,

$$(\sqrt{\sigma_n}\partial_x w_n) \rightharpoonup \sqrt{\sigma}\partial_x w \qquad \text{in w} - L^2_{t,x}. \tag{2.44}$$

Then we have

$$\|w(t)\|^2_{L^2_x} \leq \underline{\lim}\ \|w_n(t)\|^2_{L^2_x}$$

and

$$\int_0^t \int_{\mathbb{R}} \sigma(\partial_x w)^2 dxds \leq \underline{\lim} \int_0^t \int_{\mathbb{R}} \sigma_n(\partial_x w_n)^2 dxds.$$

Those two inequalities and (2.43) allow us to state that

$$\left(\|w_n(t)\|_{L^2_x}\right) \to \|w(t)\|_{L^2_x}$$

and

$$\left(\int_0^t \int_{\mathbb{R}} \sigma_n(\partial_x w_n)^2 dxds\right) \to \int_0^t \int_{\mathbb{R}} \sigma(\partial_x w)^2 dxds,$$

which implies, if we choose $t = T$ and with (2.44), that

$$(\sqrt{\sigma}_n \partial_x w_n) \to \sqrt{\sigma}\partial_x w \text{ in } L^2_{t,x}, \tag{2.45}$$

and then, with (2.11), $(\partial_x w_n) \to \partial_x w$ in $L^2_{t,x}$. Thanks to (2.43), for any t,

$$\begin{aligned} &\frac{1}{2}\left|\|w_n(t)\|^2_{L^2_x} - \|w(t)\|^2_{L^2_x}\right| \\ &\qquad \leq \|\varepsilon_n\|_{L^\infty} + \int_0^T \int_{\mathbb{R}} |\sigma(\partial_x w)^2 - \sigma_n(\partial_x w_n)^2| dxds, \end{aligned}$$

which does not depend on t and tends to 0 when $n \to \infty$ because of (2.45).

Thus we have the uniform convergence of $(\|w_n\|_{L^2_x})$ to $\|w\|_{L^2_x}$ in $C([0,T])$ and the convergence of (w_n) to w in $C_t(\text{w-}L^2_x)$, and we already know that $w \in C_t(L^2_x)$.

Therefore, we can state that $(w_n) \to w$ in $C_t(L^2_x)$, and the proof of the property (S) is finally completed. □

Remark. The solutions w_n to (2.13)–(2.14) are smooth when the coefficients in these two equations are smooth too [12].

Proof of (2.15). Let $S \in C^2(\mathbb{R})$ such that $S'' \in L^\infty$. We use (S) to prove the renormalized equation for any b, σ, and w. We consider (b_n), (σ_n), and (w_n) satisfying the property of stability (S) for the problem (2.8)–(2.9). We notice that (2.15) is obvious for smooth quantities.

Since S is C^2 with S'' bounded, it is clear that $(S''(w_n))$ tends to $S''(w)$ a.e., that $(S'(w_n))$ converges to $S'(w)$ in $C_t(L^2_{\mathrm{loc}_x})$ and that $(S(w_n))$ tends to $S(w)$ in $C_t(L^1_{\mathrm{loc}_x})$. It is then easy to let n go to $+\infty$ in the smooth version of (2.15). □

Proof of (2.16). We need here auxiliary functions $(\varphi_\alpha)_{\alpha>0}$ that are defined as follows for each $\alpha > 0$:

$$\varphi_\alpha(y) = \begin{cases} 0 & \text{if} \quad 0 \le y \le \alpha, \\ (y-\alpha)/2 + \alpha/2\pi \ \sin(\pi y/\alpha) & \text{if} \quad \alpha \le y \le 2\alpha, \\ y - 3\alpha/2 & \text{if} \quad y \ge 2\alpha, \end{cases}$$

and φ_α is an even function. For $\alpha > 0$, φ_α is positive, C^2, and convex, and φ''_α is bounded by 1. The sequence $(\varphi_\alpha)_{\alpha>0}$ uniformly tends to the absolute value when α goes to 0. We also notice, for $y \in \mathbb{R}$,

$$0 \le \varphi_\alpha(y) \le |y|, \tag{2.46}$$

$$0 \le \varphi_\alpha(y) \le \frac{\pi}{4}\frac{y^2}{\alpha}, \tag{2.47}$$

$$|\varphi_\alpha(y) - y\varphi'_\alpha(y)| \le (3/2 + 1/2\pi)\alpha = \kappa\alpha. \tag{2.48}$$

Since $w \in C_t(L^2_x)$, (2.46) and (2.47) ensure that $\varphi_\alpha(w) \in C_t(L^1_x) \cap C_t(L^2_x)$.

Remembering that φ_α is convex, i.e., $\varphi''_\alpha \ge 0$, the renormalized equation (2.15) for $S = \varphi_\alpha$ becomes an inequation in $\mathcal{D}'_{t,x}$; that is,

$$\partial_t\varphi_\alpha(w) + \partial_x(b\varphi_\alpha(w)) - \partial_x(\sigma\varphi'_\alpha(w)\partial_x w) \le \partial_x b(\varphi_\alpha(w) - w\varphi'_\alpha(w)). \tag{2.49}$$

Using a standard truncation result, we integrate (2.49):

$$\frac{d}{dt}\left(\int_{\mathbb{R}} \varphi_\alpha(w(t))dx\right) \le \kappa\alpha \int_{\mathbb{R}} |\partial_x b(t,x)|dx \qquad \text{in } \mathcal{D}'_t. \tag{2.50}$$

Hence, we can integrate (2.50) between 0 and $t \in]0,T]$ and get

$$\int_{\mathbb{R}} \varphi_\alpha(w(t))dx - \int_{\mathbb{R}} \varphi_\alpha(w_0)dx \le \kappa\alpha\sqrt{T}\|\partial_x b\|_{L^2_t(L^1_x)} = K\alpha, \tag{2.51}$$

where K is a constant (i.e., independent on t and α).

We use Fatou's lemma for $(\varphi_\alpha(w(t)))_{\alpha\in]0,1]}$, for any fixed t and find that $|w(t)| \in L^1_x$ and

$$\|w(t)\|_{L^1_x} \le \underline{\lim}\left(\int_{\mathbb{R}} \varphi_\alpha(w_0)dx + K\alpha\right) = \|w_0\|_{L^1}, \qquad 0 \le t \le T. \tag{2.52}$$

We must now prove that $w \in C_t(L^1_x)$.

LEMMA 4. *Let us consider $\gamma_1 : \mathbb{R} \to [0,1]$ a C^∞ even function, increasing on $[0,+\infty[$, such that*

$$0 \le x \le 1 \Rightarrow \gamma_1(x) = 0,$$
$$x \ge 2 \Rightarrow \gamma_1(x) = 1,$$

and, for $r \ge 1$, $\gamma_r : \mathbb{R} \to [0,1]$, $x \mapsto \gamma_1(x/r)$. Then

$$\int_{\mathbb{R}} |w(t,x)|\gamma_r(x)dx \le \int_{|x|\ge r} |w_0(x)|dx + \frac{K_0}{\sqrt{r}}, \qquad 0 \le t \le T, \tag{2.53}$$

where

$$K_0 = \lambda(b_M + \sigma_M \|\partial_x w\|_{L^2_{t,x}}), \tag{2.54}$$

and λ is a constant depending only on T, γ_1', and w_0.

Proof. We prove (2.53)–(2.54) by integrating the equation satisfied by $\varphi_\alpha(w)\gamma_r$ and using Fatou's lemma when $\alpha \to 0^+$. □

Finally, we prove that $w \in C_t(L^1_x)$ by using the fact that $w \in C_t(L^2_x)$ and Lemma 4. □

Proof of (2.17). We know that $w(t)$ belongs to $H^1_x(\mathbb{R}) \subset C_{0_x}(\mathbb{R})$ (continuous injection). By interpolation, and using (2.16), there exists a constant C_1 such that

$$\|w(t)\|_{L^\infty_x} \le C_1 \|w_0\|_{L^1}^{1/3} \|\partial_x w(t)\|_{L^2_x}^{2/3} \qquad \text{a.e. } t,$$

and we finally find (2.17). □

The proof of Proposition 3 is now completed.

Proof of Proposition 4. We first prove that, for any $t \in [0,T]$,

$$\int_{\mathbb{R}} w(t,x)dx = 0. \tag{2.55}$$

The solution w built in Proposition 3 satisfies (2.8). Using a standard truncation result, we get, in $\mathcal{D}'_t$,

$$\frac{d}{dt}\left(\int_{\mathbb{R}} w(t,x)dx\right) = 0. \tag{2.56}$$

Since, by (2.16), $t \mapsto \int_{\mathbb{R}} w(t)dx$ is continuous on $[0,T]$, we can state that

$$\int_{\mathbb{R}} w(t)dx = \int_{\mathbb{R}} w_0 dx = 0, \qquad 0 \le t \le T.$$

We are now able to define u by (2.18) and then (2.19) is obvious.

We still have to prove (2.4). It is clear that both bw and $\sigma\partial_x w$ belong to $L^2_{t,x}$. Moreover, $\partial_t w \in L^2_t(H^{-1}_x)$ implies that $\partial_t u \in L^2_{t,x}$. Hence $\mathcal{E}u := \partial_t u + bw - \sigma\partial_x w \in L^2_{t,x}$. But $\partial_x \mathcal{E}u = 0$. The two previous remarks imply (2.4).

That ends the proof of Proposition 4. □

3. Fixed point theorem. Let T, $\varepsilon > 0$. We consider the system

$$\partial_t \rho + \partial_x(\rho \tilde{u}) = 0 \qquad \text{in }]0,T[\times\mathbb{R}, \tag{3.1}$$

$$\partial_t u + \tilde{u}\partial_x u = \varepsilon \frac{\partial^2_{xx} u}{\rho} \qquad \text{in }]0,T[\times\mathbb{R}. \tag{3.2}$$

The solutions $\rho > 0$ and u must satisfy initial conditions

$$u(0,.) = u_0 \qquad \text{and} \qquad \rho(0,.) = \rho_0, \tag{3.3}$$

where $\rho_0 > 0$ and u_0 verify the following conditions:

$$\rho_0 \in L^\infty(\mathbb{R}), \qquad \frac{1}{\rho_0} \in L^\infty(\mathbb{R}), \qquad \partial_x \rho_0 \in L^1 \cap L^\infty(\mathbb{R}), \tag{3.4}$$

$$u_0 \in L^2(\mathbb{R}), \qquad \partial_x u_0 \in L^1 \cap L^2(\mathbb{R}). \tag{3.5}$$

The function $\tilde{u}$ is given and satisfies the same initial condition as u. It will be made more precise later.

3.1. Schäfer's fixed point. We use the following result which is a corollary of Schäfer's fixed point theorem [8, pp. 280–281].

THEOREM 6 (Schäfer). *Let E be a Banach space, L a closed convex set containing 0, and $Q : L \to L$ a continuous operator on L satisfying the following:*

(i) *for any closed ball B' of E centered at 0, $\overline{Q(B' \cap L)}$ is compact,*

(ii) *the set $\{x \in E \ \mid \ \exists\ \theta \in [0,1] \ \ x = \theta\ Q(x)\}$ is bounded.*

Then Q has a fixed point in L.

From now on, E will denote the space defined in (1.7), and we take

$$L = \{u \in E \ \mid \ \|u\|_{L^\infty_{t,x}} \le \|\partial_x u_0\|_{L^1}\},$$

which is a closed convex set containing 0, and

$$\mathcal{S} = \{v \in L \ \mid \ \exists\ \theta \in [0,1] \ \ v = \theta\ Q(v)\}.$$

The image $Q\tilde{u}$ of an element $\tilde{u}$ of L is obtained as follows. We solve (3.1) with initial condition ρ_0, and find $\rho > 0$ satisfying (1.8). The uniqueness of ρ follows from Lemma 2.4.1 in [2]. Next we solve the linear equation (3.2) with initial condition u_0 thanks to the results of section 2 (with $b = \tilde{u}$ and $\sigma = \varepsilon/\rho$). Finally, $Q\tilde{u} = u$, the solution of (3.2).

Remark. The image $u = Q\tilde{u}$ of $\tilde{u} \in L$ is also an element of L. First $u \in C_t(C_{0_x})$ implies that $u \in L^2_t(C_{0_x})$. Next, by (2.7), we obtain that $\partial_x u = w \in L^2_t(C_{0_x})$. The fact that $\partial_x u \in L^2_t(L^1_x)$ is a consequence of (2.16). Thus $u \in E$. It is an element of L thanks to estimate (2.19).

3.2. Properties of u. We suppose that some properties of the solution ρ to (3.1) with initial condition ρ_0 are well known [2], [6].

PROPOSITION 7. *Assuming that $\tilde{u} \in L$ and u_0 satisfies* (3.5)*, the solution u to* (3.2) *with initial condition u_0 belongs to $C_t(L^2_x)$.*

Proof. We notice that $\partial_t u \in L^2_{t,x}$, because of (3.2), and then the result is clear thanks to the injection $H^1_t(L^2_x) \subset C_t(L^2_x)$. □

PROPOSITION 8. *We consider $\tilde{u} \in L$, ρ_0, u_0 satisfying* (3.4)–(3.5)*, and let ρ and u be solutions to* (3.1)–(3.3)*. Then we have, for $S \in C^2(\mathbb{R})$, the renormalized equation*

$$\partial_t(\rho S(u)) + \partial_x(\rho \tilde{u} S(u)) - \varepsilon \partial_x(S'(u)\partial_x u) = -\varepsilon S''(u)(\partial_x u)^2. \tag{3.6}$$

Proof. Let us use property (S). We assume that ρ_{0_n}, u_{0_n}, and $\tilde{u}_n$ are smooth and verify the same assumptions as ρ_0, u_0, and $\tilde{u}$. It is easy to prove (3.6) for u_n solution to (3.2) with $\tilde{u}_n$ and ρ_n instead of $\tilde{u}$ and ρ. Since $(w_n) \to w$ in $C_t(L^2_x) \cap L^2_t(H^1_x)$, by (3.2), we can see that $(\partial_t u_n) \to \partial_t u$ in $L^2_{t,x}$. Then, with the convergence of (u_{0_n}) to u_0 in $L^2(\mathbb{R})$ and Proposition 7, we can state that $(u_n) \to u$ in $C_t(L^2_x)$. That property is sufficient to justify $n \to \infty$ in the renormalized equation depending on n. Hence (3.6) is proved. ∎

Remark. We can here prove that $\|u\|_{L^\infty_{t,x}} \leq \|u_0\|_{L^\infty}$ using (3.6) for adequate functions S. For more details on the proof, see section 7 where a similar proof is given. Note that u_0 necessarily belongs to L^∞ because of (2.16) and (2.18).

Remark. Propositions 7 and 8 and section 2 imply the smoothness (1.9) of the solution u to (1.1)–(1.2).

In the next three sections, we prove that Q satisfies conditions (i) and (ii) of Theorem 6.

4. Compactness of Q. In this section, we fix $R > 0$, and denote by $\tilde{u}$ any function in $L \cap B'(R)$.

First of all, we notice, thanks to [2] or [6], that a maximum principle applied to (3.1) gives, for a.e. t, x, knowing $\tilde{u} \in B'(R)$,

$$0 < a_0 \mathrm{e}^{-R\sqrt{T}} \leq \frac{1}{\rho(t,x)} \leq a_0' \mathrm{e}^{R\sqrt{T}}, \qquad \text{a.e. } (t,x), \tag{4.1}$$

where $a_0 = (\max \rho_0)^{-1}$ and $a_0' = (\min \rho_0)^{-1}$.

We immediately obtain a uniform estimate on $\|w(t)\|_{L^2_x}$, i.e., the following proposition.

PROPOSITION 9. *The following estimate holds:*

$$\|w(t)\|^2_{L^2_x} \leq \|w_0\|^2_{L^2} e^{R\sqrt{T}}, \qquad 0 \leq t \leq T. \tag{4.2}$$

Proof. From (2.22) and (2.25) with $b = \tilde{u}$ and $\sigma = \varepsilon/\rho$, we get (4.2). ∎

4.1. Compactness of $\partial_x u$ in $L^2_t(L^1_x)$.

PROPOSITION 10. *There exists a strictly positive constant C_1 only depending on R, ε, and initial data such that*

$$\|\partial_x w\|_{L^2_{t,x}} \leq C_1. \tag{4.3}$$

Proof. We use (2.23) and (2.26). ∎

PROPOSITION 11. *There exists a strictly positive constant C_2 depending only on R, ε, and initial data such that*

$$\|\partial_t w\|_{L^2_t(H^{-1}_x)} \leq C_2. \tag{4.4}$$

Proof. We use (2.24) and (2.27). ∎

A standard compactness result ensures from (4.3) and (4.4) the compactness of w in $L^2_t(L^1_{\mathrm{loc}_x})$. It only remains to prove that the compactness also holds in $L^2_t(L^1_x)$.

LEMMA 5. *With the notations of Lemma 4, there exists a strictly positive constant C_3 depending only on R, ε, and initial data such that, for any $r \geq 1$ and any $0 \leq t \leq T$,*

$$\int_{\mathbb{R}} |w(t)| \gamma_r dx \leq \int_{|x| \geq r} |w_0| dx + \frac{C_3}{\sqrt{r}}. \tag{4.5}$$

Proof. We only have to use Lemma 4. ∎

Lemma 5 ensures that $\sup_{t\in[0,T]}\int_{\mathbb{R}}|w(t)|\gamma_r dx$ tends to 0 when $r\to+\infty$ since the estimate (4.5) involves only R, ε, and initial data, and then, thanks to a standard compactness result, we can state the compactness of w in $L^2_t(L^1_x)$.

4.2. Compactness of $\partial_x u$ in $L^2_t(C_{0_\infty})$. Let $(\theta_n)_{n\in\mathbb{N}}$ a mollifying sequence of functions of $x\in\mathbb{R}$. We know that $w\in L^2_{t,x}$ and $\partial_t w\in L^2_t(H^{-1}_x)$. This implies, for each n, that $\theta_n *_x w\in H^1_t(C^\infty_x)$. Moreover, we have the locally compact injection $H^1_t(C^\infty_x)\subset L^2_t(C_{0_x})$. But we also have

$$\begin{aligned}|(\theta_n * w)(t,x)-w(t,x)| &\le \left(\int_{-1/n}^{1/n}\sqrt{|h|}\theta_n(h)dh\right)\|\partial_x w(t)\|_{L^2_x}\\ &\le \frac{\|\partial_x w(t)\|_{L^2_x}}{\sqrt{n}}.\end{aligned}$$

Hence, thanks to (4.3),

$$\|\theta_n *_x w - w\|_{L^2_t(C_{0_x})}\le\frac{C_1}{\sqrt{n}},$$

which proves that $(\theta_n *_x w)$ tends to w in $L^2_t(C_{0_x})$. We need the following lemma and a standard compactness result to end the proof.

LEMMA 6. *With the notations of Lemma* 4, *there exists a strictly positive constant* C_4 *depending only on* R, ε, *and initial data such that, for any* $r\ge 1$,

$$\|w\gamma_r\|_{L^2_t(C_{0_x})}\le C_4\ \|w\gamma_r\|^{1/3}_{L^2_t(L^1_x)}. \tag{4.6}$$

Proof. We prove (4.6) thanks to an interpolation result and (4.2)–(4.3). ∎

4.3. Compactness of u in $L^2_t(C_{0_\infty})$. The operator $L^2_t(L^1_x)\to L^2_t(C_{0_x})$ defined by $w\mapsto u$ is continuous, so it maps a compact set into another one. Thanks to subsection 4.1, we can state the compactness of u in $L^2_t(C_{0_x})$.

5. Continuity of Q. We prove that Q is sequentially continuous. Let $(\tilde u_n)_{n\in\mathbb{N}}\subset L$ converging to $\tilde u$ in E. As a converging sequence, $(\tilde u_n)$ is bounded in E. Let us denote $R=1+\sup\|\tilde u_n\|_E>0$.

The previous section ensures that there exists an element f of E such that, up to a subsequence, $(u_n)=(Q\tilde u_n)\to f$ in E. Let us prove that $g=\partial_x f$ satisfies the following properties:

$$\partial_t g+\partial_x(\tilde u g)=\varepsilon\partial_x\left(\frac{\partial_x g}{\rho}\right)\qquad\text{in }\mathcal{D}'_{t,x}, \tag{5.1}$$

$$g(0,.)=\partial_x u_0, \tag{5.2}$$

$$\partial_t g\in L^2_t(H^{-1}_x), \tag{5.3}$$

$$g\in L^2_t(H^1_x), \tag{5.4}$$

$$g\in C_t(L^2_x). \tag{5.5}$$

We then get $g=\partial_x(Q\tilde u)$, as a consequence of the uniqueness in Theorem 5. We successively check the properties (5.1)–(5.5).

Proof of (5.1). First, $(u_n)\to f$ in E, so $(\partial_t u_n)\rightharpoonup\partial_t f$ in $\mathcal{D}'$. Next, $(u_n)\to f$ in E implies that $(\partial_x u_n)\to\partial_x f=g$ in $L^2_t(L^1_x)$, and $(\tilde u_n)$ converges to $\tilde u$ in E, so the product sequence $(\tilde u_n\partial_x u_n)$ tends to $\tilde u g$ in $\mathcal{D}'$.

For the last term, we first notice that $(1/\rho_n) \to 1/\rho$ a.e. (a consequence of a result in [6]), and we still have (4.1). Besides, by (2.29) with $b = \tilde{u}_n$ and $\sigma = \varepsilon/\rho_n$, it is clear that $(\partial^2_{xx} u_n)$ is uniformly bounded in $L^2_{t,x}$ with respect to n. Since $(u_n) \to f$ in E and consequently in $\mathcal{D}'$, we can state that $(\partial^2_{xx} u_n)$ weakly converges in $L^2_{t,x}$ to $\partial^2_{xx} f = \partial_x g$ in w-$L^2_{t,x}$. Hence the product sequence $(\partial^2_{xx} u_n/\rho_n)$ tends to $\partial^2_{xx} f/\rho$ in $\mathcal{D}'$.

Since $\partial_t u_n + \tilde{u}_n \partial_x u_n = \varepsilon \partial^2_{xx} u_n/\rho_n$, $n \to \infty$ gives (5.1) after derivation. □

Proof of (5.2). It is obvious since $\partial_x(Q\tilde{u}_n)(0,.) = \partial_x u_0$. □

Proof of (5.3)–(5.4). We have only to use estimates (4.4) and (4.2) for each w_n and let $n \to +\infty$. □

Proof of (5.5). The estimate (4.2) implies that (w_n) is bounded in $L^\infty_t(L^2_x)$. Hence, up to a subsequence, (w_n) converges in w*-L^∞_t(w-L^2_x). But we already know that (w_n) goes to g in $\mathcal{D}'_{t,x}$. This implies that the whole sequence (w_n) converges to g in w*-L^∞_t(w-L^2_x) and then $g \in L^\infty_t(L^2_x)$. Since g verifies (5.1)–(5.2), $g \in C_t(\mathcal{D}'_x)$. Then we can state that $g \in C_t$(w-$L^2_{t,x}$) and we only have to apply Lemma 3 to end the proof. □

Now we can use uniqueness in Theorem 5 and obtain that $g = w$ and $f = u$. Thus u is the limit in E of the subsequence $(Q\tilde{u}_n)$. In fact, every converging subsequence tends to the same limit u. That means that the entire sequence $(Q\tilde{u}_n)$ converges to u, i.e., Q is continuous on E.

6. Boundedness of the set $\mathcal{S}$. We have to prove that there exists a constant β such that for any $\tilde{u} \in \mathcal{S}$,

$$\|\tilde{u}\|_{L^2_t(C_{0_x})} \leq \beta, \tag{6.1}$$

$$\|\partial_x \tilde{u}\|_{L^2_t(L^1_x)} \leq \beta, \tag{6.2}$$

$$\|\partial_x \tilde{u}\|_{L^2_t(C_{0_x})} \leq \beta. \tag{6.3}$$

By (2.16)–(2.19), and since $0 \leq \theta \leq 1$, (6.1) and (6.2) are satisfied for $\beta \geq \|w_0\|_{L^1}\sqrt{T}$. Let us prove (6.3).

First of all, we consider the system

$$\partial_t \rho + \partial_x(\rho b) = 0 \qquad \text{in }]0,T[\times\mathbb{R}, \tag{6.4}$$

$$\rho(0) = \rho_0, \tag{6.5}$$

where ρ_0 satisfies (1.5) and

$$b \in V = C_t(H^1_x) \cap L^2_t(H^2_x). \tag{6.6}$$

The existence and uniqueness of a solution to (6.4)–(6.5) in C_t(w-$L^1_{\mathrm{loc}_x}$) are proved in [2]. Moreover, the problem satisfies a property of stability [6]; for example, if $(b_n) \to b$ in V and $(\rho_{0_n}) \to \rho_0$ a.e. with a uniform bound on both $(\|\rho_{0_n}\|_{L^\infty})$ and $(1/\|\rho_{0_n}\|_{L^\infty})$, then the sequence (ρ_n) of the solutions to (6.4)–(6.5), with b_n and ρ_{0_n} instead of b and ρ_0, converges toward ρ a.e. with a uniform bound on both $(\|\rho_n\|_{L^\infty_{t,x}})$ and $(1/\|\rho_n\|_{L^\infty_{t,x}})$.

LEMMA 7. *If ρ is the solution to* (6.4)–(6.5), *ρ_0 and b satisfying* (1.5) *and* (6.6), *then both $\partial_x \ln\rho$ and $\partial_t \ln\rho$ are in $L^\infty_t(L^2_x)$, and we have, in $\mathcal{D}'_{t,x}$,*

$$\partial_t \ln\rho + b\partial_x \ln\rho = -\partial_x b \tag{6.7}$$

and

$$\partial_t(\partial_x \ln\rho) + \partial_x(b\partial_x \ln\rho) = -\partial^2_{xx} b. \tag{6.8}$$

Proof. Thanks to the property of stability for (6.4)–(6.5), we can consider $\rho_{0_n} \in C^\infty(\mathbb{R})$, $b_n, \rho_n \in C^\infty_{t,x}$. More precisely, b_n and ρ_{0_n} are regularized functions of b and ρ_0 and ρ_n is the solution to (6.4)–(6.5) with b_n and ρ_{0_n} instead of b and ρ_0. Let us set $z_n = \partial_x \ln \rho_n$. Equation (6.4) can be transformed into

$$\partial_t \ln \rho_n + b_n z_n = -\partial_x b_n, \tag{6.9}$$

$$\partial_t z_n + \partial_x (b_n z_n) = -\partial^2_{xx} b_n. \tag{6.10}$$

From (6.10), we can prove that $|z_n(t)| \in L^2_x(\mathbb{R})$ a.e. t and

$$\int_{\mathbb{R}} |z_n(t)|^2 dx \le \mathrm{e}^T \big(\|z_{0_n}\|^2_{L^2_x} + \|\partial^2_{xx} b_n\|^2_{L^2_{t,x}} \big) \exp\big(\|\partial_x b_n\|_{L^2_t(C_{0_x})} \sqrt{T} \big).$$

From (6.9), we get, for a.e. t,

$$\|\partial_t \ln \rho_n(t)\|^2_{L^2_x} \le 2\big(\|\partial_x b_n(t)\|^2_{L^2_x} + \|b_n\|^2_{C_t(H^1_x)} \|z_n(t)\|^2_{L^2_x} \big).$$

The convergences on (z_{0_n}) and (b_n) ensure that there exists $C > 0$ such that

$$\|\partial \ln \rho_n\|_{L^\infty_t(L^2_x)} \le C. \tag{6.11}$$

Besides, we know that $(\ln \rho_n) \to \ln \rho$ a.e., with a uniform bound on $\|\ln \rho_n\|_{L^\infty_{t,x}}$, so $(\partial \ln \rho_n) \rightharpoonup \partial \ln \rho$ in $\mathcal{D}'_{t,x}$. Since $(\partial \ln \rho_n)$ is bounded in $L^\infty_t(L^2_x)$, up to a subsequence,

$$(\partial \ln \rho_n) \rightharpoonup \partial \ln \rho \qquad \text{in w*-}L^\infty_t(\text{w-}L^2_x). \tag{6.12}$$

Then we can let $n \to \infty$ in (6.9) and find (6.7), (6.8), and the fact that $\partial \ln \rho \in L^\infty_t(L^2_x)$. □

Let us go back to our problem. Since $\tilde{u} = \theta u$, $\tilde{u}$ has the same smoothness as u, i.e., $\tilde{u} \in V$. Thanks to Lemma 7, we know that the derivatives $\partial \ln \rho$ are in $L^2_{t,x}$ and satisfy the following equation, in $\mathcal{D}'_{t,x}$,

$$\partial_t(\partial_x \ln \rho) + \partial_x(\tilde{u} \partial_x \ln \rho) = -\partial^2_{xx} \tilde{u}, \tag{6.13}$$

which implies that $\partial_x \ln \rho \in C_t(\mathcal{D}'_x)$. Since $\partial_x \ln \rho$ also belongs to $L^\infty_t(L^2_x)$, we can state that $\partial_x \ln \rho \in C_t(\text{w-}L^2_x)$.

We also have, in $\mathcal{D}'_{t,x}$,

$$\partial_t(\rho u) + \partial_x(\rho u \tilde{u}) = \varepsilon \partial^2_{xx} u, \tag{6.14}$$

where $\rho u \in C_t(\text{w-}L^2_x)$ thanks to the properties of both ρ and u. Using (6.13), (6.14), and $\tilde{u} = \theta u$, we get the equation, in $\mathcal{D}'_{t,x}$,

$$\partial_t F + \partial_x(\tilde{u} F) = 0, \tag{6.15}$$

after having set $F = \theta \rho u + \varepsilon \partial_x \ln \rho \in C_t(\text{w-}L^2_x)$.

We also notice that the initial condition $F_0 = \theta \rho_0 u_0 + \varepsilon \partial_x \rho_0 / \rho_0$ associated to (6.15) is in $L^1(\mathbb{R})$. Formally, by (3.6) with $S = |\cdot|$ convex, we can state that $\rho|u| \in L^\infty_t(L^1_x)$ and

$$\int_{\mathbb{R}} \rho(t)|u(t)| dx \le \int_{\mathbb{R}} \rho_0 |u_0|, \qquad 0 \le t \le T, \tag{6.16}$$

by (1.5) and (1.6). This formal computation can be justified with $S = \varphi_\alpha$. Besides, since $|\partial_x \rho_0/\rho_0| \in L^1$ too, we notice that $F_0 \in L^1$. Since, by [6], $F \in C_t(L^1_x)$ and $\|F(t)\|_{L^1_x} \leq \|F_0\|_{L^1}$, $0 \leq t \leq T$, consequently, for any t, $\partial_x \ln \rho(t) \in L^1_x$, and

$$\varepsilon \int_{\mathbb{R}} |\partial_x \ln \rho(t)| dx \leq 2 \big\| \rho_0 |u_0| \big\|_{L^1} + \varepsilon \|\partial_x \ln \rho_0\|_{L^1} = \varepsilon K_2'. \tag{6.17}$$

Using (2.16), (2.19), and (6.17), we get, for any t,

$$\int_{\mathbb{R}} |\partial_t \ln \rho(t)| dx \leq \|\partial_x u_0\|_{L^1}(1 + K_2') = K_1', \tag{6.18}$$

and then $\partial_t \ln \rho \in L^\infty_t(L^1_x)$. We can now write, for any t and a.e. x,

$$\ln \rho(t,x) = \ln \rho_0(x) + \int_0^t \partial_t \ln \rho(s,x) ds.$$

By (6.18), it is clear that $\int_0^t \partial_t \ln \rho(s,x) ds \in L^\infty_t(L^1_x)$. Let us fix $t \in [0,T]$. There exists a sequence $(x_m(t))$ which goes to $-\infty$ when $m \to +\infty$ such that

$$\int_0^t \partial_t \ln \rho(s, x_m(t)) ds \overset{m\infty}{\longrightarrow} 0. \tag{6.19}$$

But we can also write that

$$\ln \rho(t,x) = \lim_{x \to -\infty} \ln \rho(t,x) + \int_{-\infty}^x \partial_x \ln \rho(t,y) dy, \qquad x \in \mathbb{R}. \tag{6.20}$$

Using (6.20), (6.19), and letting $m \to \infty$, we find

$$\lim_{x \to -\infty} \ln \rho(t,x) = \lim_{m\infty} \ln \rho_0(x_m(t)).$$

Consequently, we find that, by (6.20) and (6.17),

$$\|\ln \rho\|_{L^\infty_{t,x}} \leq \|\ln \rho_0\|_{L^\infty} + K_2'.$$

Then there exists $K > 0$ which does not depend on θ such that

$$\frac{1}{\rho(t,x)} \geq K \qquad \text{a.e. } (t,x). \tag{6.21}$$

Moreover, by interpolation, we get, for a.e. t,

$$\|\partial_x \tilde{u}(t)\|_{C_{0_x}} \leq \big(C \|w_0\|^{1/3}_{L^1}\big) \|\partial_x w(t)\|^{2/3}_{L^2_x}, \tag{6.22}$$

by (2.16), and

$$\|w(t)\|_{L^2_x} \leq C \|w(t)\|^{2/3}_{L^1_x} \|\partial_x w(t)\|^{1/3}_{L^2_x} \leq \big(C \|w_0\|^{2/3}_{L^1}\big) \|\partial_x w(t)\|^{1/3}_{L^2_x}. \tag{6.23}$$

By (2.30) with $b = \tilde{u}$ and $\sigma = \varepsilon/\rho$, we have

$$2\varepsilon \int_0^T \int_{\mathbb{R}} \frac{(\partial_x w)^2}{\rho} dx dt \leq \|w_0\|^2_{L^2} + \int_0^T \|\partial_x \tilde{u}(t)\|_{C_{0_x}} \|w(t)\|^2_{L^2_x} dt. \tag{6.24}$$

With (6.21), (6.22), and (6.23), (6.24) becomes

$$2\varepsilon K\|\partial_x w\|^2_{L^2_{t,x}} \le \|w_0\|^2_{L^2} + CT^{1/3}\|\partial_x w\|^{4/3}_{L^2_{t,x}}.$$

Since $2 > 4/3$ in the previous equation, we can find an estimate on $\|\partial_x w\|_{L^2_{t,x}}$ that does not depend on θ, i.e.,

$$\exists\, C > 0 \qquad \|\partial_x w\|_{L^2_{t,x}} \le C. \tag{6.25}$$

The estimates (6.22) and (6.25) then imply (6.3) by Hölder's inequality.

We have checked all the conditions needed in Theorem 6 (the renormalized equation on u is easily obtained in the same way as that of $w = \partial_x u$). The fixed point of Q has all the properties of $\tilde{u}$ and u. Theorem 1 is proved, except (1.12). Note that the renormalized equation (1.10) on u is easily obtained in the same way as that of $w = \partial_x u$.

7. Upper estimate on $\partial_x u$. Let us set $v = (At+1)w$, which obviously has the same smoothness as w and satisfies

$$\partial_t v + \partial_x(uv) - \varepsilon\partial_x\left(\frac{\partial_x v}{\rho}\right) = \frac{Av}{At+1}.$$

First, we can easily prove that $(v-A)_+ \in L^2_t(H^1_x) \cap C_t(L^2_x)$. Next, as for (2.15), we notice that w satisfies the following renormalized equation in $\mathcal{D}'_{t,x}$:

$$\begin{aligned} &\partial_t S(v) + \partial_x(uS(v)) - \varepsilon\partial_x\left(\frac{S'(v)\partial_x v}{\rho}\right) \\ &= \frac{v}{At+1}\big[(A-v)S'(v) + S(v)\big] - \varepsilon S''(v)(\partial_x v)^2/\rho, \end{aligned} \tag{7.1}$$

where S is C^2 and S'' is bounded.

Let us consider, for $\alpha > 0$, the auxiliary function Φ_α defined as

$$\Phi_\alpha(y) = \begin{cases} 0 & \text{if} \quad y \le A, \\ \varphi_\alpha(y-A) & \text{if} \quad y \ge A, \end{cases}$$

where φ_α is the function we defined in section 2. For $\alpha > 0$, Φ_α is clearly positive, C^2, convex, and Φ''_α is bounded by 1. Moreover, $(\Phi_\alpha)_{\alpha\in]0,1]}$ uniformly tends to $(\cdot - A)_+$ in $C(\mathbb{R})$.

From the properties (2.46), (2.47), and (2.48) of φ_α, we can also state that

$$0 \le \Phi_\alpha(y) \le (y-A)_+, \tag{7.2}$$

$$0 \le \Phi_\alpha(y) \le \pi/(4\alpha)(y-A)^2_+, \tag{7.3}$$

$$|\Phi_\alpha(y) - (y-A)\Phi'_\alpha(y)| \le \kappa\alpha. \tag{7.4}$$

Since $(v-A)_+ \in C_t(L^2_x)$, (7.2) and (7.3) ensure that $\Phi_\alpha(v)$ is in both $C_t(L^1_x)$ and $C_t(L^2_x)$.

In (7.1), we take $S = \Phi_\alpha$ which is C^2 convex. Using a standard truncation result and noticing that $v_0 = w_0 \le A$ a.e., we get

$$\int_{\mathbb{R}} \Phi_\alpha(v(t))dx \le K\alpha, \qquad 0 \le t \le T,$$

where K is a constant which does not depend on α and t. We now use Fatou's lemma with fixed t for $(\Phi_\alpha(v(t)))_{\alpha\in]0,1]}$ and we obtain that $(v(t)-A)_+ \in L^1_x$ and

$$\int_{\mathbb{R}} (v(t)-A)_+ dx \leq 0.$$

Hence we have (1.12).

8. Inviscid limit. This section is concerned about the proof of Theorem 2.

8.1. The viscous backward problem. Equation (1.16) contains no viscosity term so that we can directly use Definition 2 of duality solutions. Indeed, L^∞ distributional solutions are duality solutions [2]. Next, we associate to (1.17) the following backward problem

$$\partial_t p^\varepsilon + u^\varepsilon \partial_x p^\varepsilon + \varepsilon \frac{\partial^2_{xx} p^\varepsilon}{\rho^\varepsilon} = 0, \tag{8.1}$$

$$p^\varepsilon(T,.) = p_T, \tag{8.2}$$

where $p_T \in \mathrm{Lip}(\mathbb{R})$ does not depend on ε and $\partial_x p_T \in L^1(\mathbb{R})$.

We set $\pi^\varepsilon = \partial_x p^\varepsilon$. If we write the system satisfied by $p^\varepsilon(T-t,x)$, we find a forward problem which can be solved thanks to Theorem 5. Hence, we can state that, since $\pi_T \in L^1$ and $p_T \in \mathrm{Lip}$, there exists a unique solution $\pi^\varepsilon \in C_t(L^2_x) \cap L^2_t(H^1_x)$, in the sense of distributions, to

$$\partial_t \pi^\varepsilon + \partial_x(u^\varepsilon \pi^\varepsilon) + \varepsilon \partial_x \left(\frac{\partial_x \pi^\varepsilon}{\rho^\varepsilon} \right) = 0, \tag{8.3}$$

$$\pi^\varepsilon(T,.) = \pi_T. \tag{8.4}$$

Then we can define p^ε by

$$p^\varepsilon(t,x) = \int_{-\infty}^{x} \pi^\varepsilon(t,y) dy + \lim_{-\infty} p_T, \tag{8.5}$$

because we can prove that $\pi^\varepsilon \in C_t(L^1_x)$ and $\frac{d}{dt}\int_{\mathbb{R}} \pi^\varepsilon(t) dx = 0$. Then we easily obtain that $p^\varepsilon \in L^\infty_{t,x}$.

Next we have to verify that p^ε defined by (8.5) satisfies (8.1). If we set $A(p^\varepsilon) = \partial_t p^\varepsilon + u^\varepsilon \partial_x p^\varepsilon + \varepsilon \partial^2_{xx} p^\varepsilon / \rho^\varepsilon$, we notice that $\partial_x[A(p^\varepsilon)] = 0$ in $\mathcal{D}'_{t,x}$ and that $A(p^\varepsilon) \in L^2_{t,x}$, which ensures that $A(p^\varepsilon) = 0$. Hence we have proved the first part of the following proposition.

PROPOSITION 12. *There exists a unique solution p^ε in the sense of distributions to* (8.1)–(8.2) *such that $\pi^\varepsilon \in C(]0,T]; L^2_x) \cap L^2(]0,T[; H^1_x)$ and*

$$\|p^\varepsilon\|_{L^\infty_{t,x}} \leq \|p_T\|_{L^\infty} + \|\partial_x p_T\|_{L^1}.$$

It satisfies the following a priori estimate:

$$|\partial_x p^\varepsilon(t,x)| \leq \frac{T}{t} \|\partial_x p_T\|_{L^\infty} \qquad \text{a.e. } t,x. \tag{8.6}$$

Moreover, if p_T is monotone, then p^ε has the same monotonicity in x.

Remark. This monotonicity property is compatible to the notion of reversible solutions for the inviscid problem.

Proof. In this proof, we will not use the notations with ε. Let us set $\nu(t,.) = \pi(t,.)t/T$, which satisfies the following renormalized equation, where N is an arbitrary constant and $S \in C^2(\mathbb{R})$ is such that S'' is bounded:

$$\begin{aligned}(8.7)\qquad &\partial_t S(\nu) + \partial_x(u\nu + \varepsilon\partial_x\nu/\rho)\\ &= (\nu - Nt\partial_x u)S'(\nu)/t + \partial_x u(S(\nu) - \nu S'(\nu)) + \varepsilon(\partial_x\nu)^2 S''(\nu)/\rho.\end{aligned}$$

First, we can choose $N = \max(\operatorname{ess\,sup}\pi_T, 0)$ and, at least formally, $S(y) = (y-N)_+$. Knowing the smoothness of ν, we obtain, in the same way as in section 7, that $(\nu - N)_+$ has the same smoothness as ν. Then, noticing that the chosen function S is convex, we get, in $\mathcal{D}'_{t,x}$,

$$\partial_t(\nu - N)_+ \geq (\nu - Nt\partial_x u)\frac{\mathbf{1}_{\nu>N}}{t} = 0,$$

since we know that $\nu \leq N$ and $\partial_x u \leq 1/t$. That gives us, for any $0 < t \leq T$,

$$\int_{\mathbb{R}} (\nu(t) - N)_+ dx \leq \int_{\mathbb{R}} (\nu_T - N)_+ dx = 0,$$

and finally

$$(8.8)\qquad \pi(t) \leq \frac{T}{t}N \leq \frac{T}{t}\|\pi_T\|_{L^\infty_{t,x}} \qquad \text{a.e.}$$

Remark. The previous formal computation can easily be justified with approximate functions as done in section 7.

Note that, if we choose a nonincreasing final datum p_T in (8.2), N becomes equal to 0 in (8.8) and we obtain the required monotonicity property.

To complete the proof, we can set, for example, $N = -\max(\operatorname{ess\,sup}(-\pi_T), 0)$ and choose $S(y) = (y-N)_-$. Then it is easy to prove the other side of the desired estimate and end the proof. ∎

We now want to justify the asymptotics $\varepsilon \to 0$ in the backward problem. Let us choose $\tau \in]0, T[$. We need the following lemma.

LEMMA 8. *The following a priori estimate holds:*

$$\|\pi^\varepsilon(t)\|_{L^2_x} \leq \frac{T}{t}\|\pi_T\|_{L^2} \qquad \textit{a.e. } t \in [\tau, T].$$

Proof. From (8.3), we can compute that

$$(8.9)\qquad \frac{d}{dt}\|\pi^\varepsilon(t)\|^2_{L^2_x} + \int_{\mathbb{R}} \partial_x u^\varepsilon(t)\pi^\varepsilon(t)^2 dx - 2\varepsilon\int_{\mathbb{R}} \frac{(\partial_x\pi^\varepsilon(t))^2}{\rho^\varepsilon(t)}dx = 0,$$

which implies, thanks to Gronwall's lemma and (1.12) and after having noticed that the term $\int_{\mathbb{R}} \frac{(\partial_x\pi^\varepsilon)^2}{\rho^\varepsilon}dx$ is nonnegative, the desired estimate. ∎

We are now able to prove the following asymptotics result.

PROPOSITION 13. *There exists a subsequence of the solutions (p^ε) to (8.1)–(8.2) which converges in $C([\tau, T]; C_x)$ toward the solution p to (1.14) with final data p_T, where u is the weak* limit in $L^\infty_{t,x}$ of a subsequence of (u^ε).*

Proof. Let us prove that the last term $(\varepsilon\partial_x\pi^\varepsilon/\rho^\varepsilon)$ goes to 0 when $\varepsilon \to 0$ in, say, $L^2_{t,x}$. From (8.9) integrated between τ and T, we get

$$\varepsilon\int_\tau^T\int_{\mathbb{R}} \frac{(\partial_x\pi^\varepsilon)^2}{\rho^\varepsilon}dxdt \leq \frac{1}{2}\|\pi_T\|^2_{L^2} + \int_\tau^T \frac{\|\pi^\varepsilon(t)\|^2_{L^2_x}}{2t}dt \leq C,$$

where C is a constant which does not depend on ε, because of Lemma 8.

Moreover, taking (1.12) into account and thanks to the assumptions about ρ_0^ε and u_0^ε, a maximum principle applied to (1.16) gives

$$\frac{1}{\rho^\varepsilon(t,x)} \leq \frac{(1+\max(\text{ess sup}\,\partial_x u_0^\varepsilon,0))}{\rho_0^\varepsilon(x)} \leq \frac{C}{\sqrt{\varepsilon}} \qquad \text{a.e.,}$$

where C does not depend on ε.

Hence, thanks to Cauchy–Schwartz's inequality,

$$\varepsilon^2 \int_\tau^T \int_{\mathbb{R}} \frac{(\partial_x \pi^\varepsilon)^2}{(\rho^\varepsilon)^2} dx dt \leq C\sqrt{\varepsilon}, \tag{8.10}$$

and then $(\varepsilon \partial_x \pi^\varepsilon/\rho^\varepsilon)$ tends to 0 in $L^2_{t,x}$ when $\varepsilon \to 0$. We must now prove the asymptotics on $(u^\varepsilon \partial_x p^\varepsilon)$. In fact, we study the sequences $(p^\varepsilon u^\varepsilon)$ and $(\partial_x u^\varepsilon p^\varepsilon)$.

From Lemma 8, we know that $(\partial_x p^\varepsilon)$ is uniformly bounded in $L^\infty_t(L^2_x)$, and, from (8.1), (8.10), and Lemma 8, we easily find that $(\partial_t p^\varepsilon)$ is also uniformly bounded in $L^2_{t,x}$. Consequently, we have, by interpolation, a uniform estimate on (p^ε) in both $L^\infty_t(C^{0,1/2}_x)$ and $C^{0,1/2}_t(L^2_x)$. Hence, up to a subsequence, $(p^\varepsilon) \to p$ in $C_{t,x}$. Since (u^ε) is uniformly bounded in $L^\infty_{t,x}$, there exists a subsequence of (u^ε) converging in w*-$L^\infty_{t,x}$ to a limit u, and then $(p^\varepsilon u^\varepsilon)$ converges to pu in $\mathcal{D}'_{t,x}$.

To study the term $p^\varepsilon \partial_x u^\varepsilon$, we set $\mu^\varepsilon = 1/t - \partial_x u^\varepsilon$. Thanks to (1.12), we can state that μ^ε is a nonnegative measure. We want to prove that (μ^ε) is bounded in the sense of measures. In fact, we first notice that it is bounded in the sense of distributions. More precisely, if $\varphi \in \mathcal{D}_{t,x}$,

$$|\langle \mu^\varepsilon, \varphi \rangle_{\mathcal{D}',\mathcal{D}}| \leq \frac{|K|}{\tau} \|\varphi\|_{L^\infty_{t,x}} + C|K| \|\partial_x \varphi\|_{L^\infty_{t,x}},$$

where K is the compact support of φ and $|K|$ its Lebesgue measure. Then, since μ^ε is nonnegative, we easily find that (μ^ε) is bounded in the sense of measures. This implies that $(\partial_x u^\varepsilon)$ is uniformly measure-bounded too. Then there exists a subsequence of $(\partial_x u^\varepsilon)$ that is convergent in the sense of measures. But since $(p^\varepsilon) \to p$ in $C_{t,x}$, we find that $(p^\varepsilon \partial_x u^\varepsilon) \rightharpoonup p \partial_x u$ in $\mathcal{D}'_{t,x}$.

Finally we have proved that p satisfies

$$\partial_t p + u \partial_x p = \partial_t p + \partial_x (pu) - p \partial_x u = 0.$$

The previous equation proves that $p \in C([\tau,T]; \text{w-}L^2_x)$. Consequently, the final condition is satisfied too. □

8.2. Proof of Theorem 2. Now that we know (p^ε) converges towards p, we have to justify that (q^ε) is convergent too. First of all, thanks to (1.20) and (6.16), it is clear that (q^ε) is uniformly bounded in the sense of measures; thus, up to a subsequence, (q^ε) is convergent.

Besides, knowing the smoothness of p^ε, ρ^ε, and u^ε, we have only to prove that

$$\frac{d}{dt}\left(\int_{\mathbb{R}} p^\varepsilon \rho^\varepsilon u^\varepsilon dx \right) = 0 \qquad \text{in } [\tau,T], \tag{8.11}$$

as a generalization of the notion of duality solutions.

Each term in (8.1) is in $L^2_{t,x}$, so we can multiply it by $q^\varepsilon \in L^2_{t,x}$ and get

$$q^\varepsilon \partial_t p^\varepsilon + q^\varepsilon u^\varepsilon \partial_x p^\varepsilon + \varepsilon u^\varepsilon \partial^2_{xx} p^\varepsilon = 0. \tag{8.12}$$

On the other hand, thanks to Lemma 7, we know that $\partial\rho^\varepsilon \in L^\infty_t(L^2_x)$. Since $u^\varepsilon, p^\varepsilon \in L^\infty_{t,x}$, and $\partial_t u^\varepsilon, \partial_x u^\varepsilon, \partial^2_{xx} u^\varepsilon \in L^2_{t,x}$, we can justify the multiplication of (1.17) by p^ε and get

$$p^\varepsilon \partial_t(q^\varepsilon) + p^\varepsilon \partial_x(q^\varepsilon u^\varepsilon) - \varepsilon p^\varepsilon \partial^2_{xx} u^\varepsilon = 0. \tag{8.13}$$

We sum (8.12) and (8.13), and we get

$$\partial_t(p^\varepsilon q^\varepsilon) + \partial_x(p^\varepsilon q^\varepsilon u^\varepsilon) + \varepsilon \partial_x(u^\varepsilon \partial_x p^\varepsilon - p^\varepsilon \partial_x u^\varepsilon) = 0.$$

With a standard truncation result, noticing that $p^\varepsilon q^\varepsilon, p^\varepsilon q^\varepsilon u^\varepsilon \in L^1_{t,x}$, and $u^\varepsilon \partial_x p^\varepsilon - p^\varepsilon \partial_x u^\varepsilon \in L^2_{t,x}$, we find (8.11).

Thus we have proved that a subsequence of (q^ε) converges in $C_t(\text{w*-}\mathcal{M}_{\text{loc}_x})$ to q satisfying

$$\frac{d}{dt}\int_{\mathbb{R}} p(t,x) q(t,dx) = 0 \qquad \text{in } [\tau, T] \times \mathbb{R}.$$

Thus (1.19) is satisfied in the duality sense in any interval $]t_1, t_2[$ with $0 < t_1 < t_2$. The density ρ also satisfies (1.18) in the duality sense, and it can be directly obtained by the stability of duality problems [2], which also gives the relation $q = \rho u$.

Acknowledgment. The author wishes to thank F. Bouchut for his advice and the numerous talks we had on this topic.

REFERENCES

[1] F. Bouchut, *On zero pressure gas dynamics*, in Advances in Kinetic Theory and Computing , Ser. Adv. Math. Appl. Sci. 22, World Scientific, River Edge, NJ, 1994, pp. 171–190.

[2] F. Bouchut and F. James, *One-dimensional transport equations with discontinuous coefficients*, Nonlinear Anal., 32 (1998), pp. 891–933.

[3] F. Bouchut and F. James, *Solutions en dualité pour les gaz sans pression*, C. R. Acad. Sci. Paris Sér. I Math., 326 (1998), pp. 1073–1078.

[4] F. Bouchut and F. James, *Duality solutions for pressureless gases, monotone scalar conservation laws, and uniqueness*, Comm. Partial Differential Equations, 24 (1999), pp. 2173–2189.

[5] Y. Brenier and E. Grenier, *Sticky particles and scalar conservation laws*, SIAM J. Numer. Anal., 35 (1998), pp. 2317–2328.

[6] R.J. DiPerna and P.-L. Lions, *Ordinary differential equations, transport theory and Sobolev spaces*, Invent. Math., 98 (1989), pp. 511–547.

[7] E. Weinan, Y.G. Rykov, and Y.G. Sinai, *Generalized variational principles, global weak solutions and behavior with random initial data for systems of conservation laws arising in adhesion particle dynamics*, Comm. Math. Phys., 177 (1995), pp. 349–380.

[8] D. Gilbarg and N.S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Springer-Verlag, Berlin, 1983.

[9] E. Grenier, *Existence globale pour le système des gaz sans pression*, C. R. Acad. Sci. Paris Sér. I Math., 321 (1995), pp. 171–174.

[10] D. Hoff, *The sharp form of Oleinik's entropy condition in several space dimensions*, Trans. Amer. Math. Soc., 276 (1983), pp. 707–714.

[11] D. Hoff and R. Zarnowski, *Continuous dependence in L^2 for discontinuous solutions of the viscous p-system*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 11 (1994), pp. 159–187.

[12] J.-L. Lions and E. Magenes, *Problèmes aux limites non homogènes*, Dunod, Paris, 1968.

[13] P.-L. Lions, *Mathematical Topics in Fluid Mechanics,* Vol. 2*, Compressible Models*, Clarendon Press, Oxford, UK, 1998.

[14] F. Poupaud and M. Rascle, *Measure solutions to the linear multidimensional transport equation with discontinuous coefficients*, Comm. Partial Differential Equations, 22 (1997), pp. 337–358.

[15] E. Zeidler, *Nonlinear Functional Analysis and Its Applications,* Vol. II/A. *Linear Monotone Operators*, Springer-Verlag, Berlin, 1990.

[16] Ya.B. Zeldovich, *Gravitational instability: An approximate theory for large density perturbations*, Astron. and Astrophys., 5 (1970), pp. 84–89.

SIAM J. MATH. ANAL.
Vol. 32, No. 1, pp. 194–213

EXISTENCE OF UNDERCOMPRESSIVE TRAVELING WAVES IN THIN FILM EQUATIONS*

A. L. BERTOZZI† AND M. SHEARER‡

Abstract. We consider undercompressive traveling wave solutions of the partial differential equation

$$\partial_t h + \partial_x f(h) = -\partial_x(h^3\partial_x^3 h) + D\partial_x(h^3\partial_x h),$$

when the flux function f has the nonconvex form $f(h) = h^2 - h^3$. In numerical simulations, these waves appear to play a central role in the dynamics of the PDE; they also explain unusual phenomena in experiments of driven contact lines modeled by the PDE. We prove existence of an undercompressive traveling wave solution for sufficiently small nonnegative D and nonexistence when D is sufficiently large.

Key words. undercompressive shocks, traveling waves, heteroclinic orbit, existence

AMS subject classifications. 34C37, 35L65, 35L67, 35Q35, 76D08, 76D45

PII. S0036141099350894

1. Introduction.

The partial differential equation (PDE)

$$\partial_t h + \partial_x f(h) = -\partial_x(h^3\partial_x^3 h) + D\partial_x(h^3\partial_x h) \tag{1.1}$$

describes the flow of a thin liquid film on an inclined flat surface, under the action of gravity, viscous, and surface tension forces. Parameters governing these forces, and the slope of the surface, are incorporated into the dimensionless parameter $D \geq 0$. In particular, $D = 0$ for a vertical surface. The unknown function $h = h(x,t)$ is the (dimensionless) thickness of the thin film layer.

Equation (1.1) arises from the standard lubrication approximation of the Navier–Stokes equations [BB97, BMS99, Gre78]. We consider the specific physical problem in which the film is driven by two counteracting forces, namely, gravity pulling the film down the plane, and a thermal gradient, which induces a surface tension gradient, pushing the film up the plane. The interested reader should see [BMFC98, BMS99] for a discussion of (1.1) and the dimensionless scaling. For this particular problem, the dimensionless flux function in (1.1) is

$$f(h) = h^2 - h^3. \tag{1.2}$$

Equation (1.1) results when we assume the film height is independent of an additional transverse space variable (cf. (6.3) at the end of this paper). Experimental and numerical studies of driven contact lines [THSJ89, BB97, BMFC98, JSMB98] show that traveling wave solutions of the PDE (1.1) play an important role in the motion of the film. The significance of the nonconvexity of the flux function in (1.2) is that (1.1) then admits the possibility of undercompressive traveling waves, which we discuss in detail below.

*Received by the editors February 2, 1999; accepted for publication (in revised form) September 22, 1999; published electronically June 22, 2000.
http://www.siam.org/journals/sima/32-1/35089.html

†Duke University, Department of Mathematics, Box 90320, Durham, NC 27708 (bertozzi@math.duke.edu). This author was supported by a PECASE award from the ONR.

‡Department of Mathematics, North Carolina State University, Box 8203, Raleigh, NC 27695-8203. This author was supported by NSF grant DMS-9818900 and ARO grant AG55-98-10128.

Driven contact line experiments that can be modeled by (1.1) show some unusual dynamics. First, there are experiments in which there is only one dominant driving force, corresponding to a convex flux function f. For example, $f(h) = h^2$ in the case of dominant Marangoni stress [CHTC90, KT97] or $f(h) = h^3$ in the case of gravitational stress [Hup82, THSJ89, dB92]. For such examples, the film forms a pronounced "capillary ridge" which corresponds mathematically to a nonmonotone traveling wave solution of (1.1). The ridge results from the interaction of surface tension, in the form of the fourth order diffusion on the right-hand side of (1.1), with the driving force, in the form of the convective term $(f(h))_x$. Such ridges have always been associated with instabilities of the film that lead to the formation of finger-like structures [BB97, THSJ89, dB92, JdB92, VIC98] in which h develops a growing oscillatory dependence on the transverse variable.

Secondly, it is interesting to contrast the driven contact line experiments where one force dominates with those experiments involving competing Marangoni and gravitational stresses. Early experiments [LL71] of relatively thick Marangoni-gravity driven films show a stable front with *monotone* decrease of the film profile from the bulk to the contact line. Recent experiments [Fan98, BMFC98] show that for intermediate thickness films, a capillary ridge forms but continues to broaden while the *contact line remains stable* and no fingering occurs. The model (1.1) with the nonconvex flux (1.2) has recently been used to establish that undercompressive traveling waves are responsible for the unusual ensuing dynamics of the front [MB99, BMFC98]. In these papers, the prewetted surface is modeled as a thin precursor layer, avoiding unresolved issues of how to model a propagating liquid/solid/air contact line.

The consideration of traveling waves reduces the fourth order partial differential equation to a third order ordinary differential equation (after integrating once) depending on two parameters, namely the wave speed and the downstream film thickness. In the three-dimensional phase space of the ODE, compressive waves correspond to a codimension zero intersection of the two-dimensional unstable manifold of one equilibrium with the two-dimensional stable manifold of another equilibrium. Generically, this intersection is transverse and hence structurally stable, persisting under perturbations of the equation. For convex flux functions, e.g., $f(h) = h^2$ (Burgers flux), existence of compressive waves follows either from the analysis of Kopell and Howard [KH75] or from an argument involving a Lyapunov function and the Conley index [Mic88, Ren96, BMS99].

In contrast, undercompressive waves, which only arise when the flux is nonconvex, correspond to a codimension one intersection of the one-dimensional unstable manifold of one equilibrium with the two-dimensional stable manifold of another equilibrium. This situation typically only occurs for special values of the parameters in the ODE. The analysis of such special connections is straightforward for corresponding problems in second order ODEs; the phase space is two-dimensional and the Melnikov integral gives a measure of the separation of the manifolds in question. In our situation, the phase space is three-dimensional and the argument is more difficult. Our proof of existence of the undercompressive wave uses, in a central way, a Lyapunov function for the ODE to analyze the behavior of the one-dimensional unstable manifold from the largest equilibrium of the system. We combine this analysis with a shooting argument involving both topological properties of the orbit and quantitative estimates of higher derivatives of the solution and of its turning points. The techniques presented here apply to more general nonconvex flux functions than (1.2) and may be useful in understanding other higher dimensional bifurcation problems.

Undercompressive shock waves have been found in other physically motivated models involving systems of equations with application to dynamic phase transitions in elastic solids [AK91, Jam80, She86], liquid/vapor phase transitions [Sle83, Tru87], plane magnetohydrodynamic waves [Fre97], and multiphase flow related to secondary oil recovery [IMP90, SSMPL87, IMPT92]. Moreover, undercompressive waves have been analyzed in nonconvex conservation laws, with second order dissipation and (third order) dispersion [HL97, HS98, JMS95]. The model (1.1) represents the first realization of undercompressive shocks arising in a scalar conservation law with direct connection to experiments. The fourth order nonlinear diffusion, which has its own curious properties (see [Ber98] and references therein), combined with the nonconvex flux f yields undercompressive waves.

In an earlier paper [BMS99], we identified numerically new traveling wave solutions of (1.1) for $D = 0$ that correspond to undercompressive shock wave solutions of the conservation law. In this paper, we prove the existence of undercompressive traveling waves for small $D \geq 0$. Specifically, we show that for each downstream film thickness h_+ there is an undercompressive traveling wave, provided $D \geq 0$ is not too large. On the other hand, we also show that if D is large enough, then there is no undercompressive traveling waves with right state h_+. The latter property agrees with the limit $D \longrightarrow \infty$, for which second order diffusion dominates, and the theory is classical [Smo94]. In section 2 we discuss preliminaries concerning the phase space, and in section 3 we introduce the Lyapunov function that plays a major part in making the shooting argument work. section 4 contains the proof of existence of the undercompressive waves, while section 5 is a proof of nonexistence for large enough D.

2. Preliminaries. We are interested in traveling wave solutions of the equation

$$\partial_t h + \partial_x f(h) = -\partial_x(h^3\partial_x^3 h) + D\partial_x(h^3\partial_x h), \tag{2.1}$$

with $f(h) = h^2 - h^3$, and $D \geq 0$. On long scales, solutions of (2.1) behave like solutions of the corresponding scalar conservation law

$$\partial_t h + \partial_x f(h) = 0. \tag{2.2}$$

For this equation, recall that characteristics are straight lines

$$\frac{dx}{dt} = f'(h),$$

on which h is constant. A piecewise constant function

$$h(x,t) = \begin{cases} h_- & \text{if} \quad x < st, \\ h_+ & \text{if} \quad x > st \end{cases} \tag{2.3}$$

is a *shock wave solution* (with shock speed s) if the triple h_-, h_+, s satisfies the Rankine–Hugoniot condition

$$-s(h_+ - h_-) + f(h_+) - f(h_-) = 0. \tag{2.4}$$

A shock wave is *compressive* if the characteristics on each side of the shock impinge on the shock. This property is the *Lax entropy condition:*

$$f'(h_+) < s < f'(h_-). \tag{2.5}$$

As we shall see, undercompressive waves violate the Lax entropy condition.

A traveling wave solution $h = h(\xi)$, $\xi = x - st$, of (2.1) with speed s that has far field limits

$$\lim_{\xi\to-\infty} h(\xi) = h_- \text{ and } \lim_{\xi\to\infty} h(\xi) = h_+ \tag{2.6}$$

can be thought of, on large scales, as a "viscous" form of the shock (2.3). The existence of stable traveling wave profiles of (2.1) connecting the state h_- to the state h_+ is a criterion for the admissibility of the shock (2.3) in the large scale dynamics of (2.1). We are interested in the possibility of admissible undercompressive shocks, violating (2.5).

In general, traveling waves satisfy the third order ODE

$$-s(h - h_+) + f(h) - f(h_+) = -h^3h''' + Dh^3h'. \tag{2.7}$$

(In integrating the equation once, we have assumed $h'(\xi) \to 0$ and $h'''(\xi) \to 0$ as $\xi \to \infty$.) Equation (2.7) has two parameters $h_+ \in (0, 1/3)$ and $s > 0$. Possible left states $h = h_-$ (where $h' = 0 = h'''$) are determined by (2.4), the Rankine–Hugoniot condition for shocks.

To discuss (2.7), we begin by rewriting it:

$$h''' = g(h; h_+, s) + Dh', \tag{2.8}$$

where

$$g(h; h_+, s) = -h^{-3}\left(-s(h - h_+) + f(h) - f(h_+)\right). \tag{2.9}$$

At an equilibrium of (2.9), $h = h_e$, $g(h_e, h_+, s) = 0$, and the linearized ODE $u''' = \frac{\partial g}{\partial h}(h_e; h_+, s)u + Du'$ has characteristic equation

$$\lambda^3 - D\lambda - \frac{\partial g}{\partial h}(h_e; h_+, s) = 0. \tag{2.10}$$

For $D = 0$, the three eigenvalues are simply the three cube roots of $\frac{\partial g}{\partial h}(h_e; h_+, s)$. Since $\frac{\partial g}{\partial h}(h_e; h_+, s) = -\frac{1}{h_e^3}(f'(h_e) - s)$, the sign of $\frac{\partial g}{\partial h}(h_e; h_+, s)$ at any equilibrium h_e is related to whether characteristics at h_e, traveling with speed $f'(h_e)$, are faster or slower than the speed s of the traveling wave. For $0 \le D < 3(\frac{1}{2}\frac{\partial g}{\partial h}(h_e; h_+, s))^{2/3}$, there is one real eigenvalue $\lambda(D)$ (satisfying $\lambda(0) = (\frac{\partial g}{\partial h}(h_e; h_+, s))^{1/3}$), and two complex conjugate eigenvalues $\lambda_\pm(D)$. For larger D, all three eigenvalues are real. Moreover,

$$\lambda(D) \neq 0 \quad \text{and} \quad \mathrm{sgn}\Re(\lambda_\pm(\mathrm{D})) = -\mathrm{sgn}\lambda(\mathrm{D}) \quad \text{for all} \quad \mathrm{D}.$$

To describe the structure of equilibria, we write (2.8) as a first order system:

$$\begin{aligned} h' &= v, \\ v' &= w, \\ w' &= g(h; h_+, s) + Dv. \end{aligned} \tag{2.11}$$

We have the following classification of nondegenerate equilibria $(h, v, w) = (h_e, 0, 0)$ for (2.11).

(i) If $f'(h_e) < s$, then $\frac{\partial g}{\partial h}(h_e; h_+, s) > 0$, so that $(h_e, 0, 0)$ has a one-dimensional unstable manifold and a two-dimensional stable manifold on which, for small D,

solutions spiral into the equilibrium due to the complex conjugate pair of eigenvalues with negative real part.

(ii) If $f'(h_e) > s$, then $\frac{\partial g}{\partial h}(h_e; h_+, s) < 0$, so that $(h_e, 0, 0)$ has a one-dimensional stable manifold and a two-dimensional unstable manifold on which, for small D, solutions spiral away from the equilibrium due to the complex conjugate pair of eigenvalues with positive real part.

It is convenient to label the equilibria in order of their corresponding values of h. Physically h_+ plays the role of a precursor layer in an experiment. Thus the relevant range is for h_+ small. Define $b = h_+ \in (0, 1/3)$ and let this be fixed. Treating s as a parameter, let[1] $h = h_m(s) \le h_t(s)$ be the two roots (different from h_+) of (2.4):

$$h + b - (h^2 + bh + b^2) = s$$

for s in the range

$$s_1 = f'(b) \le s \le \frac{2(f((1-b)/2) - f(b))}{(1-3b)} = s_2. \tag{2.12}$$

For brevity, we sometimes write $m = h_m(s), t = h_t(s)$. In particular (see Figure 2.1),

$$b < m < \frac{1-b}{2} < t < 1 - 2b \quad \text{if} \quad s_1 < s < s_2,$$

and

$$b = m; \quad t = 1 - 2b \quad \text{if} \quad s = s_1, \qquad m = t = (1-b)/2 \quad \text{if} \quad s = s_2.$$

Moreover, (with $h_+ = b$) the vector field (2.11) has three equilibria when $s_1 < s < s_2 : B = (b, 0, 0), M = (m, 0, 0), T = (t, 0, 0)$. From the discussion of equilibria above, we see that B and T each have a one-dimensional unstable manifold and a two-dimensional stable manifold, whereas M has a two-dimensional unstable manifold and a one-dimensional stable manifold.

The arguments of Kopell and Howard [KH75, BMS99] show that if B and M (or M and T) are sufficiently close, then there is a trajectory from M to B (or M to T, respectively). The corresponding traveling wave is necessarily compressive since $f'(b) < s < f'(m)$ (and $f'(t) < s < f'(m)$). Such trajectories lie along the intersection of the two-dimensional unstable manifold from M and the two-dimensional stable manifold from B (or T, respectively). This construction is structurally stable in that it persists under small perturbations of the vector field (for example, by changing s while keeping b fixed).

Undercompressive waves correspond to trajectories from T to B, or from B to T. These occur when the one-dimensional unstable manifold from T (from B, respectively) lies in the two-dimensional stable manifold from B (from T, respectively), a codimension one construction. The main result of section 4 is that for b fixed and for all small $D \ge 0$, there is a value of s for which there is such a trajectory from T to B. (The corresponding result from B to T follows by a symmetric argument, but is less significant physically.)

In section 5 we show that for each $b < 1/3$, and for D sufficiently large, there is no value of s for which there is an undercompressive traveling wave from T to B. This result expresses the notion that for large D, second order diffusion dominates fourth order diffusion. In the absence of fourth order diffusion, the only traveling waves are compressive.

[1]Note that the subscript t here does not denote partial derivative. It is an index to denote the specific equilibrium.

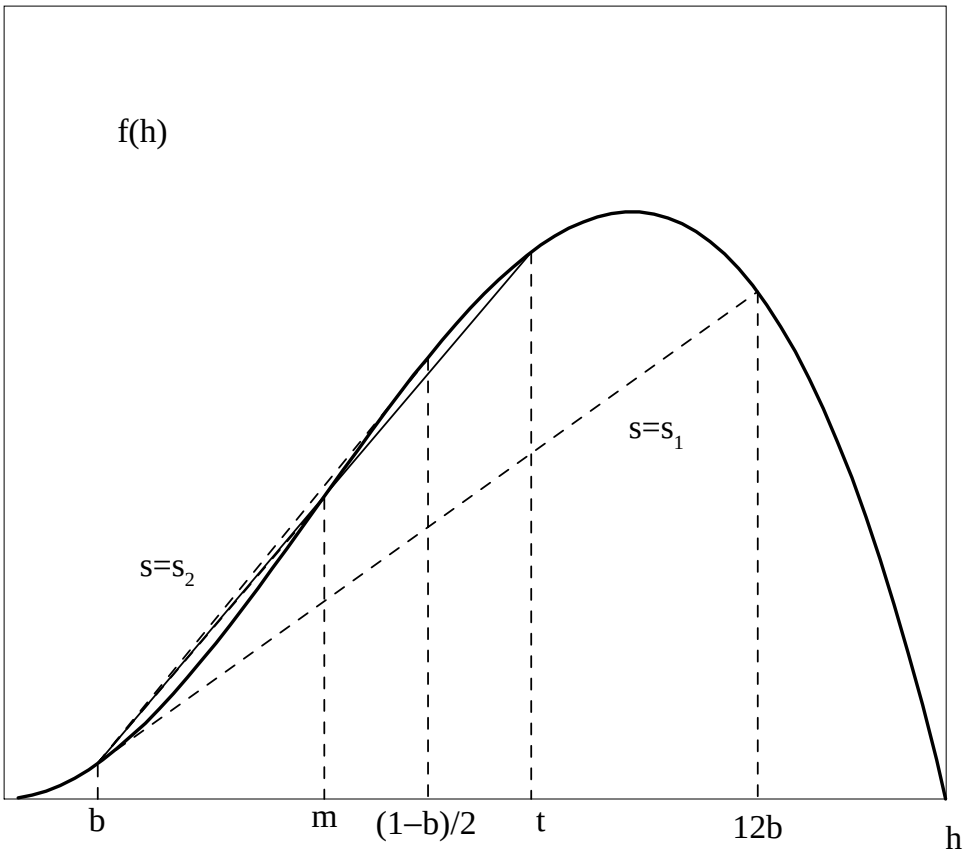


FIG. 2.1. *Flux function $f(h) = h^2 - h^3$, and chords indicating equilibria and wave speeds.*

3. The Lyapunov function. Equation (2.8) has a Lyapunov function

$$L(h) = h''h' + R(h),$$

where

$$\frac{dR}{dh}(h) = -g(h; b, s),$$

which we use extensively in the analysis of traveling waves. The equilibria B, M, T correspond to extrema b, m, t of $R(h)$, as shown in Figure 3.1.

Differentiating along a solution $h(\xi)$ and using the ODE (2.8), we find that

$$L(h)' = (h'')^2 + D(h')^2.$$

Therefore, $L(h)$ increases along trajectories. In particular, $R(h)$ increases at successive critical points of a solution $h(\xi)$ of (2.8). It follows that for any traveling wave solution connecting extrema of $R(h)$ there exist a priori upper and lower bounds for the critical

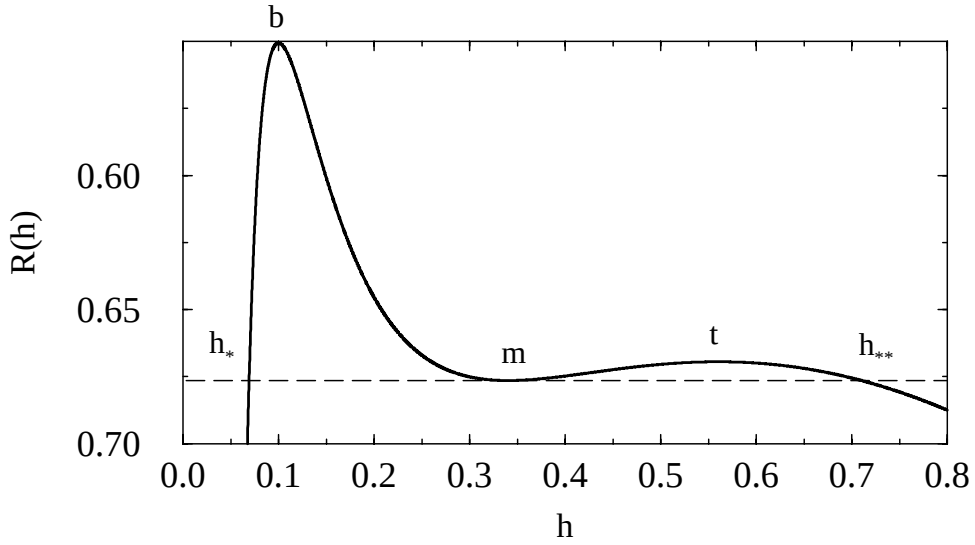


FIG. 3.1. *The function $R = -dg/dh$ in the Lyapunov function. Pictured are the three equilibria, b, m, and t, of* (2.11) *and the a priori upper and lower bounds h_{**} and h_*, defined in* (3.2)*, for a traveling wave solution.*

points h_{crit} of the traveling wave

$$h_* < h_{crit} < h_{**}, \tag{3.1}$$

where h_* and h_{**} are defined by (see Figure 3.1)

$$h_* = \min\{h : R(h) \geq R(m)\}, \quad h_{**} = \max\{h : R(h) \geq R(m)\}. \tag{3.2}$$

Note that R and hence h_*, h_{**} depend on b and s.

4. Existence of an undercompressive wave. In this section, we fix $h_+ = b < 1/3$, and consider the vector field (2.11) with s and D varying. An undercompressive wave occurs when there exists a trajectory (a *heteroclinic orbit*) from the equilibrium $T = (h_t(s), 0, 0)$ to the equilibrium $B = (b, 0, 0)$. We show that for sufficiently small D (depending on the value of b), there exists a special value of s, call it s_*, for which there is such a trajectory.

For each value of s, we consider a special solution of (2.11), corresponding to the branch of the unstable manifold from the equilibrium $(h_t(s), 0, 0)$ that initially decreases in h. Let $h_t(\xi; s)$, $-\infty < \xi$, denote the solution of (2.8) corresponding to this branch. By the stable manifold theorem and the Picard continuation theorem for ODEs, $h_t(\xi; s)$ is smooth and tangent to the unstable manifold of the linearized ODE about h_t and is determined uniquely up to translation in ξ. The goal of this section is to prove that there is an s_* for which

$$\lim_{\xi \longrightarrow \infty} h_t(\xi; s_*) = b.$$

The proof of this result was inspired in part by results from numerical simulations [BMS99, M99]. The argument is a one parameter, one direction shooting argument. In Proposition 4.1 below, we show that for s near s_1, for which b and $h_m(s)$ are

close, $h_t(\xi; s)$ decreases monotonically, reaching zero at a finite value of ξ. On the other hand, in Proposition 4.2, we show that for s near s_2, so that $h_m(s)$ and $h_t(s)$ are close, $h_t(\xi; s)$ has a minimum value above $h = b$; the trajectory then increases without bound. The trajectory we seek lies between these two extremes; its existence is established in Theorem 4.8. While this paper establishes existence of such a special shock speed, uniqueness remains an open problem. However, numerical computations [BMS99, M99] reveal the shock speed and undercompressive wave to be unique.

The first part of this argument is based on the Lyapunov function. First note that there is a value of s, call it s_l, such that the two maxima of $R(h; s)$, b and $h_t(s)$ satisfy $R(b; s) = R(h_t(s); s)$ and

$$R(b; s) < R(h_t(s)) \quad \text{for all} \quad s \quad \text{satisfying } f'(b) < s < s_l.$$

The function R has a global maximum at $h_t(s)$ for s in this range. Figure 3.1 shows a case where $s > s_l$; in this case R has a global maximum at b.

PROPOSITION 4.1. *For all s, $f'(b) < s < s_l$, $h_t(\xi; s)$ decreases monotonically to hit zero at a finite value of ξ.*

Proof. Suppose that $h_t(\xi; s)$ has a local minimum at a finite $\xi = \xi_0$. Then, necessarily $v_t(\xi; s) \equiv (h_t(\xi; s))_\xi$ is zero at ξ_0. Since the Lyapunov function increases along trajectories, we have $R(h_t(\xi_0; s) \geq R(h_t(s))$. However, this contradicts the fact that $h_t(s)$ is a global maximum of R for this range of s. Thus there can be no local minimum at finite ξ_0.

Now we show the solution decreases to hit zero at finite ξ. To see this, we note that since h is monotonically decreasing, it either hits zero at finite ξ_0 or it stays positive for all $\xi < \infty$, which means that because it is decreasing, it has a limit $h \to h_0 \geq 0$. We now show that the latter case leads to a contradiction.

First suppose that $h_0 > 0$. Then the only choices are $h_0 = b$ or $h_0 = h_m(s)$. Otherwise, (2.8) implies that $h''' - Dh'$ will remain bounded away from zero on an interval of the kind $[l_0, \infty)$ which implies that $h'' - Dh$, and hence h'' and h' become unbounded, which is a contradiction. The two equilibria, b and $h_m(s)$ are also ruled out by the properties that $L(h)$ is an increasing function of ξ, $L = R$ at equilibria, and R has a global maximum at $h = h_t(s)$. Thus we can not have that $h_0 > 0$.

Now suppose h decreases monotonically to zero in infinite time. Again, from the ODE, this implies that eventually the $h''' - Dh'$ becomes monotonically unbounded, inconsistent with h decreasing monotonically to zero.

The only choice then is for $h \to 0$ at some finite ξ. ∎

PROPOSITION 4.2. *Let $b \in (0, 1/3)$. There are numbers $D_0, \underline{s}, \overline{s}$ with $s_\ell < \underline{s} < \overline{s} < s_2$ such that for all $D \in [0, D_0]$ and all $s \in [\underline{s}, \overline{s}]$, $h_t(\xi, s)$ has a global minimum between $h_m(s)$ and b. The solution then increases without bound after reaching that minimum.*

Proof. It suffices to prove that $h_t(\xi; s)$ has a global minimum between $h_m(s)$ and b. The result then follows from the Lyapunov function and a similar argument to the first part of the proof of Proposition 4.1.

First we prove that for $D = 0$, there is a range of s, $s_u < s < s_2$ for which $h_t(\xi, s)$ has the property claimed in the proposition. Then we use a perturbation argument to prove the result for small positive D.

To show that $h_t(\xi; s)$ has such a minimum, we first estimate the trajectory at $h_m(s)$ in terms of the parameter $\rho = h_t(s) - h_m(s)$, which decreases to zero as s approaches s_2. In what follows, we consider $\rho > 0$ to be small.

LEMMA 4.3. *Let $D = 0$. Then $h_t(\xi_m; s) = h_m(s)$ for some $\xi_m < \infty$.*

Proof. To simplify notation, consider s fixed, and write $h = h(\xi)$ in place of $h = h_t(\xi, s)$. Then h has the properties

$$(h, h', h'') \longrightarrow (h_t(s), 0, 0) \quad \text{as} \quad \xi \longrightarrow -\infty,$$

and moreover

$$(h, h', h'') \sim (h_t(s), 0, 0) - Ce^{\lambda\xi}(1, \lambda, \lambda^2) \quad \text{as} \quad \xi \longrightarrow -\infty,$$

where

$$\lambda = \left(\frac{\partial g}{\partial h}(h_t(s); b, s)\right)^{1/3}$$

is the positive eigenvalue for the equilibrium $(h_t(s), 0, 0)$ for (2.11). In particular, for ξ sufficiently negative,

$$h(\xi) < h_t(s), v(\xi) = h'(\xi) < 0, w(\xi) = h''(\xi) < 0. \tag{4.1}$$

Now define the open set $O \subset R^3$ by

$$O = \{(h, v, w) : h_m(s) < h < h_t(s), v < 0\}.$$

Then $(h, v, w)(\xi) \in O$ for all $\xi < M$, for some M.

Next note that the vector field (2.11) is uniformly Lipschitz in O, since the only nonlinearity is in $g(h; b, s)$, a function of h alone (for fixed b, s) whose derivative is bounded for $h_m(s) \le h \le h_t(s)$. Therefore, the solution $(h, v, w)(\xi)$ can be continued in ξ as long as it remains in O. That is, either (a) the solution stays in O for all $\xi \in R$, or (b) the solution exits S at some finite $\xi = \xi_T$.

In case (a), the solution must approach an equilibrium as $\xi \longrightarrow \infty$. Thus, $(h, v, w) \longrightarrow (h_m(s), 0, 0)$, or $(h, v, w) \longrightarrow (h_t(s), 0, 0)$. Both possibilities are ruled out by the property that the Lyapunov function must increase along the trajectory, since $R(h_m(s))$ and $R(h_t(s))$ are both less than or equal to the value of the Lyapunov function at $\xi = -\infty$.

In case (b), $(h, v, w)(\xi_T) \in \partial O$. Thus, (i) $h'(\xi_T) = 0$, or (ii) $h(\xi_T) = h_t(s)$, or (iii) $h(\xi_T) = h_m(s)$. But in O, $h''' = g(h; b, s) < 0$, so that h'' is decreasing, hence negative, by (4.1). But this implies that h' is decreasing, and must also be strictly negative, contradicting (i) and ruling out (ii). Hence, (iii) holds, completing the proof of the lemma. □

To proceed further with the proof of Proposition 4.2, we parameterize the unstable manifold

$$\{(h, v, w)(\xi) : -\infty < \xi < \infty\}$$

by h, up to the first minimum of $h(\xi)$. That is, we consider $v = v(h) = h'(\xi)$, and we write (2.8) with $D = 0$ as a nonautonomous equation for $v(h)$:

$$v^2 v'' + v(v')^2 = g(h; b, s). \tag{4.2}$$

The solution of (4.2) we consider satisfies $v(t) = 0$, and $v'(t) > 0$. In fact, higher derivatives of v at $h = t$ can easily be determined using the Taylor series of $v(h)$ about

$h = t$. To simplify notation, we write $h_t(s) = t$, and $h_m(s) = m$. Then $\rho = t - m > 0$. Next, define a new function $G(h;\rho)$ for $\rho \geq 0, t - \rho \leq h \leq t$ by

$$G(h,\rho) = \begin{cases} g(h;b,s)/[(h-t)(h-t+\rho)] & \text{for} \quad t-\rho < h < t, \\ -\dfrac{1}{\rho}\dfrac{\partial g}{\partial h}(t-\rho;b,s) & \text{if} \quad h = t-\rho, \\ \dfrac{1}{\rho}\dfrac{\partial g}{\partial h}(t;b,s) & \text{if} \quad h = t. \end{cases}$$

Then G is as smooth as g, except at $h = t, h = t - \rho$, where, in general, G loses a derivative. For the specific g in this paper, both g and G are rational functions of h, so there is no loss of derivative. Note that $G(h,\rho) > 0 \quad \text{for} \quad t - \rho \leq h \leq t$, and $g(h;b,s) = (h-t)(h-t+\rho)G(h,\rho)$.

We scale (4.2) as follows: Write

$$h = t + \rho\theta, \quad v = \rho^{4/3} y(\theta). \tag{4.3}$$

Then (4.2) becomes

$$y^2 y'' + y y'^2 = \theta(\theta+1)G(t+\rho\theta,\rho). \tag{4.4}$$

From the Taylor series expansion of $y(\theta)$ about $\theta = 0$, where we impose the condition $y(0) = 0$, we find

$$y'(0) = G(t,\rho) = G(t,0) + O(\rho). \tag{4.5}$$

Now Lemma 4.3 implies the following.

LEMMA 4.4. *There are constants $\rho_0 > 0, 0 < \alpha < \beta$, such that for each $\rho \in (0,\rho_0)$, the solution $y(\theta), -1 \leq \theta \leq 0$, of (4.4) satisfying $y(0) = 0$, (4.5) also satisfies*

$$-\beta < y(-1) < -\alpha, \quad \alpha < y'(-1) < \beta.$$

Proof. First note that $\theta = -1$ corresponds to ξ_m in Lemma 4.3. The scaled Lyapunov function is $L(y) = y^2 y' + R(\theta)$, where

$$R(\theta) = \int_\theta^0 \eta(\eta+1)G(t+\rho\eta,\rho)d\eta < 0$$

for $-1 \leq \theta < 0$. Specifically, $\frac{d}{d\theta}L(y) = y y'^2 < 0$, so that $L(y)$ increases as θ, hence y decreases. If $y' = 0$, then $L(y) = R(\theta) < R(0) = 0 = L(0)$. Thus, $L(y)$ has not increased, which is a contradiction. Therefore, $y' > 0$ along the entire trajectory from $\theta = 0$ to $\theta = -1$. The result now follows by bounding the $O(\rho)$ term in G. ∎

From the rescaling (4.3) and the chain rule, we conclude the following.

COROLLARY 4.5. *Let $v(h)$ be the solution of (4.2) corresponding to $y(\theta)$ of Lemma 4.4. Then*

$$-\rho^{4/3}\beta < v(m) = \rho^{4/3} y(-1) < -\rho^{4/3}\alpha; \quad \rho^{1/3}\alpha < \frac{dv}{dh}(m) = \rho^{1/3}\frac{dy}{d\theta}(-1) < \rho^{1/3}\beta. \tag{4.6}$$

Now, for $b \le h \le m$, $g(h) > 0$, so that as long as $v(h) < 0$, (4.2) implies that $v''(h) > 0$. Thus v is a convex function of h whenever v is negative. We use this fact below.

To complete the proof of Proposition 4.2, we suppose that $v(h) < 0$ for $b < h < m$ and look for a contradiction. The idea is to show that for small enough ρ, by estimating v and $v'(h)$ over half this interval, when we integrate (4.2) the integral of g remains bounded away from zero, while the integral of the left-hand side approaches zero.

Note that for all $b < h < m$,

$$0 > v(h) = v(m) - \int_h^m v'(\eta)d\eta \ge v(m) - (m-h)v'(m)$$

(since $v'(\eta) \le v'(m)$ for $\eta \le m$). Therefore,

$$|v(h)| \le |v(m)| + (m-h)K_1\rho^{1/3} \le K\rho^{1/3}, \quad h_M \le h \le m,$$

for some $K > 0$ independent of ρ.

Also, the convexity of v and inequality (4.6) imply that

$$K\rho^{1/3} > v'(m) > v'(h) \ge \frac{v(h) - v(b)}{h - b} \tag{4.7}$$

for all $b < h < m$. Now consider $h_M = (m + b)/2$. The above inequalities imply that

$$|v(h_M)| \le K\rho^{1/3}, \qquad |v'(h_M)| \le K_1\rho^{1/3},$$

where in the second inequality we use (4.7), the bounds on v, and the fact that h_m is not close to b.

Now we integrate (4.2) from h_M to m, integrating the left-hand side by parts:

$$v(m)^2v'(m) - v(h_M)^2v'(h_M) - \int_{h_M}^m v(h)v'(h)^2dh = \int_{h_M}^m g(h)dh.$$

But the left-hand side is order ρ, while the right-hand side is order one, as $\rho \longrightarrow 0$. This contradiction implies that $v(h) = 0$ for some $h \in (b, m)$, for each $\rho > 0$ sufficiently small.

To summarize, we have so far shown that for $D = 0$, there is a range $s_u < s < s_2$ for which the unstable manifold from t decreases to a global minimum between m and b and then increases without bound. To continue the proof of Proposition 4.2, we need to establish the same behavior for small $D > 0$. Since the unstable manifold from t depends continuously on D, away from $s = s_2$ (at $s = s_2$, two equilibria coincide, so the unstable manifold degenerates), there is $D_0 > 0$ and two values of s, say $s_u < \underline{s} < \overline{s} < s_2$ such that for $0 \le D \le D_0$, $\underline{s} \le s \le \overline{s}$, the unstable manifold from t has $h_\xi = v$ changing sign for h between m and b. It then follows from the Lyapunov function argument used previously that the solution $h(\xi; s)$ has a global minimum between m and b.

Finally we note that the solution increases without bound after the local minimum between m and b. This is because, like the preceding arguments based on the Lyapunov function, the solution cannot have a local maximum after hitting this minimum and cannot asymptote to either the fixed point m or b. This completes the proof of Proposition 4.2. □

We now define two distinguished values of h. Let

$$\overline{h} = \max_{s_1 \le s \le s_2} h_{**}(s), \quad \underline{h} = \min_{s_1 \le s \le s_2} h_*(s), \tag{4.8}$$

where $h_* = h_*(s), h_{**} = h_{**}(s)$ are given by (3.2).

LEMMA 4.6. *For all $s \in (s_1, s_2)$, the trajectory $h_t(\xi, s)$ crosses the boundary of the set $\overline{h} < h < \underline{h}$ at most once, either by increasing h above $\underline{h}$ or by decreasing h below $\overline{h}$. In the former case, the solution increases without bound after it leaves this set and in the latter case, the solution hits zero at finite ξ.*

Proof. Suppose the trajectory $h_t(\xi, s)$ crosses the lower boundary $\underline{h}$. Then it is impossible for the solution to turn around. If it did, there would be a local minimum at a value $h_{\min} < \underline{h}$, which by the definition of $\underline{h}$ in (4.8) violates the Lyapunov condition (3.1). Likewise if $h_t(\xi, s)$ crosses the upper boundary $\overline{h}$ it cannot turn around because this would again violate (3.1). ∎

Now define

$$S = \{s \in (s_1, s_2) | h_t(\xi, s) \text{ increases above } \overline{h} \text{ for some finite } \xi\}.$$

For all D satisfying the conditions of Proposition 4.2 we know that S is not empty; it contains at least one interval near s_2. Also, from Proposition 4.1, we know that S does not contain any $s < s_l$. Thus for all D satisfying the conditions of Proposition 4.2, the following special value of s is well defined:

$$s_* = \inf\{S\}. \tag{4.9}$$

Clearly $s_* \ge s_l$.

LEMMA 4.7. *Let D satisfy the conditions of Proposition* 4.2 *and s_* be defined as in* (4.9). *Then the trajectory $h_t(\xi; s_*)$ remains bounded between $\overline{h}$ and $\underline{h}$ and can be continued in this range for all $\xi < \infty$.*

Proof. First we note that $h_t(\xi; s_*)$ stays below $\overline{h}$. Suppose it crosses $\overline{h}$ at finite $\xi = \xi_0$ (i.e., $s_* \in S$). Since solutions of (2.11) have continuous dependence on the parameter s, there then exists an $\epsilon > 0$ so that $s_* - \epsilon' \in S$ for all $0 < \epsilon' < \epsilon$. This contradicts the fact that $s_* = \inf S$. Now we show that $h_t(\xi; s_*)$ stays above $\underline{h}$ for all ξ. Suppose it does not. Then there exists a value ξ_0 at which $h_t(\xi; s_*)$ crosses $\underline{h}$. Again, since solutions of (2.11) have continuous dependence on the parameter s, there exists an $\epsilon > 0$ so that for all $\epsilon > \epsilon' > 0$, $h_t(\xi; s_* + \epsilon')$ crosses the lower bound $\underline{h}$. Hence $s_* + \epsilon' \notin S$. However, this contradicts the fact that s_* is the infimum.

Thus the trajectory $h_t(\xi; s_*)$ is guaranteed to stay between $\underline{h}$ and $\overline{h}$. We need to show that the trajectory can be continued for all time. As in the proof of Lemma 4.3, we can do this by using the continuation part of the Picard theorem for ODEs, provided we can show uniform Lipschitz continuity of $(v, w, g(h))$ as a function of (h, v, w). Since the solution is guaranteed to have h bounded between $\underline{h}$ and $\overline{h}$, by the form of g, Lipschitz continuity is guaranteed for the third component. Moreover, the other two terms are linear in v and w, so that uniform Lipschitz continuity is guaranteed for (h, u, v) in the set $[\underline{h}, \overline{h}] \times R \times R$. The solution can thus be continued for all $\xi < \infty$. ∎

THEOREM 4.8. *Given D satisfying the conditions of the statement of Proposition* 4.2, *and s_* defined in* (4.9), *the unstable manifold $h_t(\xi, s_*)$ (with s_* defined as above) connects the equilibrium $h_t(s_*)$ to the equilibrium b and hence describes an undercompressive wave.*

Proof. We have that $h_t(\xi, s_*)$ is bounded. For ease of notation below, we denote this trajectory simply by $h(\xi)$.

Case 1: There exists a finite $\xi_{\max}$ above which $h(\xi)$ has no extrema, i.e., it is monotone increasing or decreasing. Since h is bounded, it is convergent to a limit as $\xi \to \infty$. That limit must be an equilibrium. If not, then h''' is uniformly bounded away from zero on a semi-infinite line $[\xi_0, \infty)$ and we can show this causes h, h', and h'' to become unbounded. That equilibrium has to be either m, b, or t. However, by comparing the values of the Lyapunov function, we see that the only choice is b, since $R(m) < R(h_t)$, the Lyapunov function can be shown to initially increase by comparing the solution with the predicted linear theory. Note that although this case does imply that $h(\xi) \to b$ as $\xi \to \infty$, this case is not the expected scenario. Note that when D is small, the stable manifold of b has two complex conjugate eigenvalues so that we would expect a trajectory on it to spiral in to b, i.e., we expect such a solution to have an infinite number of local extrema as $\xi \to \infty$.

Case 2: There exists a set of points $X \in R$ where h has an extremum and hence $h_\xi = 0$, and $\sup\{X\} = \infty$.

First note that such points are isolated. This is because the a priori upper and lower bounds on the solution and the fact that it satisfies (2.8) imply that $h(\xi)$ is a global real analytic function, and hence if there is a cluster point for $h_\xi = 0$, then h_ξ must be identically zero, which is clearly not the case. Thus the set X must be a countable set ξ_i and $\xi_i \to \infty$ as $i \to \infty$.

Denote by h_i the value $h(\xi_i)$. Let us suppose without loss of generality that ξ_i are local minima for i odd and maxima for i even. From the Lyapunov function, we see that all extrema satisfy $R(h(\xi_i)) > R(h_t)$. Moreover $R(h_i)$ is an increasing sequence that is also bounded, so it converges to a value R_1. Furthermore, all minima must lie below b. This is because if, say, h_k lies above b, then since $R(h)$ is monotone decreasing on the set $R(h) > R(h_t)$, $h > b$, then $R(h_{k+1}) < R(h_k)$ because h_{k+1} is a local maximum. However, this contradicts the fact that $R(h_i)$ is increasing. Likewise, a similar argument shows that the local maxima all lie above b.

By (3.1) and (4.8), the solution lies between $\underline{h}$ and $\overline{h}$. The proof follows provided we can show that h' and h'' approach zero as $\xi \to \infty$.

To do this, we make some explicit estimates, using (2.8) and the Lyapunov function. First note that since $R(h_i)$ is increasing, the h_i oscillate around $h = b$: $h_i < b$ at a min and $h_{i+1} > b$ at a max. Therefore, there are two convergent subsequences, $h_{2n+1} \to h_1$, $h_{2n} \to h_2$, with $h_2 - h_1 \geq 0$.

Now note that for each ξ_i,

$$R(h_i) - R(h_t) = \int_{-\infty}^{\xi_i} h_{\xi\xi}^2 + D \int_{-\infty}^{\xi_i} h_\xi^2.$$

Taking the limit as $i \to \infty$ and recalling that $D \geq 0$ gives

$$\int_{-\infty}^{\infty} h_{\xi\xi}^2 d\xi \leq D \int_{-\infty}^{\infty} h_\xi^2 d\xi + \int_{-\infty}^{\infty} h_{\xi\xi}^2 d\xi = R_1 - R(h_t) \leq R(b) - R(h_t) < \infty.$$

We now invoke the following interpolation inequality [Tay96, p. 9];

$$\|h_\xi\|_{L^4(R)} \leq C \|h\|_{L^\infty}^{1/2} \|h_{\xi\xi}\|_{L^2(R)}^{1/2}.$$

This means that since h is uniformly bounded and $h_{\xi\xi}$ is bounded in $L^2(R)$ that h_ξ is bounded in $L^4(R)$.

On $[\xi_i, \xi_{i+1}]$, h_ξ has the fixed sign $(-1)^{i+1}$. Now choose $\beta_i \in [\xi_i, \xi_{i+1}]$ so that $|h_\xi|$ attains a maximum on this interval at β_i. Compute

$$\begin{aligned}
|h_\xi^3(\beta_{i+1}) - h_\xi^3(\beta_i)| &= |h_\xi(\beta_{i+1})|^3 + |h_\xi(\beta_i)|^3 \\
&= 3\left|\int_{\beta_i}^{\beta_{i+1}} h_\xi^2 h_{\xi\xi} d\xi\right| \\
&\le 3\int_{\beta_i}^{\beta_{i+1}} |h_\xi|^2 |h_{\xi\xi}| d\xi \\
&\le 3\left[\int_{\beta_i}^{\beta_{i+1}} |h_\xi|^4\right]^{1/2}\left[\int_{\beta_i}^{\beta_{i+1}} |h_{\xi\xi}|^2 d\xi\right]^{1/2} \\
&\le 3\epsilon_i\delta_i,
\end{aligned}$$

where

$$\delta_i = \left[\int_{\beta_i}^{\beta_{i+1}} |h_\xi|^4\right]^{1/2}, \quad \epsilon_i = \left[\int_{\beta_i}^{\beta_{i+1}} |h_{\xi\xi}|^2 d\xi\right]^{1/2},$$

and where $\sum_i \epsilon_i^2$ and $\sum_i \delta_i^2$ are both finite. Thus $\epsilon_i\delta_i \to 0$ as $i \to \infty$. By the choice of β, this also implies that $|h_\xi|^3$ and hence $|h_\xi|$ goes to zero as $\xi \to \infty$. Note that this also implies that h_ξ is uniformly bounded on R.

We now show that $h_{\xi\xi}$ is uniformly bounded independent of ξ. Since h solves the ODE (2.11), and since h is uniformly bounded between $\underline{h}$ and $\overline{h}$, we have that $h_{\xi\xi\xi} - Dh_\xi$, and hence $h_{\xi\xi\xi}$ is uniformly bounded. Thus for any ξ,

$$|h_{\xi\xi}^3(\xi)| = 3|\int_{-\infty}^{\xi} h_{\xi\xi}^2 h_{\xi\xi\xi} d\xi| \le C\|h_{\xi\xi}\|_{L^2}^2 < \infty.$$

Finally note that the Lyapunov function is $h_\xi h_{\xi\xi} + R(h)$. Since it is increasing, and the product $h_\xi h_{\xi\xi}$ goes to zero as $\xi \to \infty$, then $R(h(\xi))$ approaches a constant. The infinite sequence of alternating max and mins implies that that constant has to be $R(b)$.

Finally, since the trajectory $\{(h, h', h'')(\xi) : -\infty < \xi < \infty\}$ is bounded, and there are no periodic orbits, the trajectory must approach an equilibrium as $\xi \longrightarrow \infty$. The equilibrium is necessarily b, since this is the only equilibrium with $R(h) > R(t)$. This completes the proof of Theorem 4.8 ∎

5. Nonexistence of undercompressive waves for large D. In this section, we show that for each $b < 1/3$, and for D sufficiently large, there are no undercompressive traveling wave solutions having $h = b$ as the downstream height. This is formulated precisely in the following theorem, in which, as in the previous section, we fix $h_+ = b$, and consider the vector fields (2.11) to be parameterized by s and D.

THEOREM 5.1. *Let $b \in (0, 1/3)$. Then there is $D_1 > 0$ such that for $D > D_1$ and $s_1 < s < s_2$, there is no orbit from the equilibrium $(h_t(s), 0, 0)$ to the equilibrium $(b, 0, 0)$.*

Proof. We need to show that the unstable manifold from $h_t(s)$ never connects to the fixed point at b.

Recall the ODE is

$$\begin{aligned} h' &= v, \\ v' &= w, \\ w' &= g(h) + Dv, \end{aligned} \tag{5.1}$$

where

$$g(h) = -\frac{f(h) - sh - f(b) + sb}{h^3}.$$

Some of the results in the preceding section apply to this case, in particular Proposition 4.1 and Lemma 4.6. Using these two results, we now show that for sufficiently large D (depending on b) for all s in the region that we are interested in, h_t decreases monotonically to zero, in which case it can never connect to the fixed point at b.

To see that this is true, first note that the linearization of (5.1) near the fixed point $h_t(s)$ yields eigenvalues that satisfy the equation

$$\lambda^3 - \lambda D - g'(h_t) = 0.$$

Since $g'(h_t) \geq 0$, for small D there is one positive real root and two complex roots with negative real part. For sufficiently large D there is one positive real root $\lambda_p \sim \sqrt{D}$ and two real negative roots $\lambda_1 \sim -g'(h_t)/D$ and $\lambda_2 \sim -\sqrt{D}$.

First we note that the Lyapunov function guarantees that the branch of the unstable manifold from h_t that initially increases can not turn around to connect to b. This is because if the solution turns around, it must have a local maximum above h_t; however, the function $R(h)$ decreases monotonically above h_t.

Consider now the branch of the unstable manifold from h_t that initially decreases. We show that for D sufficiently large, this branch decreases to zero at finite ξ.

To linear order the solution looks like

$$h_t(\xi; s) = h_t - e^{\lambda_p \xi} \tag{5.2}$$

for ξ very negative. Also, to linear order,

$$v \sim \lambda_p (h - h_t),$$

and as long as $h > \underline{h}$ (defined in (4.8)) we have an a priori bound for $g(h)$. In particular, we can choose D large enough so that the ODE (5.1) is dominated by the linear behavior (i.e., $g = 0$) of the ODE while $h > \underline{h}$. However, the linear behavior simply has that h decreases monotonically like (5.2). So for D large enough, the solution should decreases monotonically until it hits $\underline{h}$. However, we know that once it hits $\underline{h}$ it continues to decrease by Proposition 4.1.

We now make the above argument rigorous. Introducing the new variables

$$Q = \frac{v}{h - h_t}, \quad P = \frac{w}{h - h_t},$$

the ODE (5.1) is transformed to the system

$$\begin{aligned} Q' &= P - Q^2, \\ P' &= B(h(\xi)) + DQ - QP, \end{aligned} \tag{5.3}$$

$$\text{where } B(h) = \frac{g(h)}{h - h_t}.$$

Note that since g vanishes at h_t, B is bounded and approaches g' at h_t. Also, for $\underline{h} < h < h_t$, $|B|$ is bounded independent of D. Call this bound $M(b)$. This system has a fixed point for $h = h_t$ that corresponds to the positive eigenvalue λ_p above, $Q = \lambda_p$, $P = \lambda_p^2$. Now consider the rectangle

$$R_D = \{(Q,P) | \sqrt{D}/2 < Q < 2\sqrt{D} \text{ and } D/2 < P < 2D\}. \tag{5.4}$$

Choose D to satisfy $D > (4M)^{2/3}$. Then as long as $\underline{h} < h < h_t$, on the boundary of R_D, the vector field in (5.3) points into R_D, which means that the solution remains in R_D. This gives a lower bound on $Q = \frac{h_\xi}{h-h_t}$,

$$\frac{h_\xi}{h - h_t} \geq K > 0,$$

which implies $h_\xi < K(h - h_t)$ so that $h(\xi)$ decreases exponentially: for all $\xi > \xi_0$, $h(\xi) - h_t \leq (h(\xi_0) - h_t)e^{K(\xi - \xi_0)}$ provided $\underline{h} < h(\xi') < h_t$ for all $\xi' < \xi_0$. By the stable manifold theorem, we know there exists such an ξ_0 where $\underline{h} < h(\xi_0) < h_t$. This is sufficient to guarantee that $h(\xi)$ hits $\underline{h}$ at a finite value of ξ. □

6. Summary and conclusions. We have considered traveling wave solutions $h(x - st)$ of the PDE

$$\partial_t h + \partial_x(h^2 - h^3) = -\partial_x(h^3 \partial_x^3 h) + D\partial_x(h^3 \partial_x h). \tag{6.1}$$

Recent numerical experiments [BMS99, M99] show that certain jump initial data give rise to undercompressive structures, in which the leading part of the structure is an undercompressive traveling wave, connecting states h_- to h_+, for which the speed s of the wave violates the Lax entropy condition

$$f'(h_+) < s < f'(h_-).$$

For a fixed value of h_+, the numerics show a special value of h_- for which an undercompressive waves exists when the parameter D in (6.1) is small. Likewise, for large D, the numerics show that undercompressive waves do not exist. In this paper we presented rigorous proofs of both of these numerical observations.

Traveling waves satisfy a third order autonomous ODE in which the downstream thickness h_+ and the wave speed s appear as parameters. For each $h_+ = b < 1/3$, there is a range of s for which the ODE has three (hyperbolic) equilibria, B, M, and T. M has a two-dimensional unstable manifold while B and T have two-dimensional stable manifolds. Compressive waves are heteroclinic orbits from M to either B or T, codimension zero intersections of a two-dimensional stable manifold from one fixed point with a two-dimensional unstable manifold from another fixed point. Such intersections are structurally stable to perturbations and exist for a range of the parameter s. In contrast, undercompressive waves are heteroclinic connections from either T (or B) to B (or T, respectively). The situation that corresponds to the physical problem of interest is the existence of a wave from T to B.

Our analysis relies heavily on the existence of a Lyapunov function for the ODE. It follows directly from the Lyapunov function that there is a range of s (where M is close to B) for which a branch of the unstable manifold from T decreases monotonically to zero. We then consider a range of s for which M is very close to T and rescale the ODE using the distance from M to T as a scaling parameter. By analyzing the

rescaled equation for $D = 0$, we are able to show that whenever M is sufficiently close to T, the initially decreasing branch of the unstable manifold from T has a global minimum (in h) between m and b. Moreover, a perturbation argument shows this property for $D \geq 0$ and small, provided M is not too close to T. We then proceed with an argument that shows that there is an intermediate value of the parameter s, so that M is neither very close to B or to T for which the unstable manifold from T must connect to B. This part of the proof is largely topological, but it includes some explicit estimates on higher derivates of the solution along the unstable manifold from T in order to guarantee that it stays bounded and hence connects to B.

In the last section of the paper, we show that for large values of D, regardless of the speed s of the wave, the unstable manifold from T never connects to B. The result is that there can never exist an undercompressive wave. The proof follows from making a change of variables in the ODE to show that the linear system dominates the dynamics along the unstable manifold, until the solution gets so small in h that it must hit zero at finite ξ.

We mention some related papers discussing third order (ODE) travelling waves that exist only for special parameter values or wave speeds. The paper of Grinfeld [Gri89] deals with travelling waves for Korteweg capillary regularization of a van der Waals fluid and uses Conley index theory to prove existence. The paper [BHP96] deals with traveling waves in the compressive case $f(h) = h^3$ but with a different form of degenerate diffusion. They prove existence of waves with a sharp contact line (h goes to zero) using a two directional shooting method. It would be interesting to see if the methods of these papers also apply to the problem presented here.

It is interesting to note that our arguments extend, with slight modifications, to the case of linear diffusion:

$$\partial_t h + \partial_x(f(h)) = D\partial_x^2 h - \partial_x^4 h. \tag{6.2}$$

In fact, it is the fourth order diffusion that produces the undercompressive shocks. Numerical simulation of (6.2) shows that similar structures occur in this case. The main difference between (2.1) and (6.2) is that the degeneracy in the diffusion in (2.1), in particular in the fourth order term, causes some singular behavior to occur for very small values of h. Numerical computations of the traveling waves for $D = 0$ show that as $b \to 0$, the value of s for which the undercompressive wave occurs approaches $s = 0$ while the value of t approaches $t = 1$. For very small values of b, jump initial data corresponding to very weak Lax shocks evolve to a solution of (6.1) with two shocks, with the special undercompressive wave as the leading shock. Since this undercompressive wave connects $t \sim 1$ to $b \sim 0$, we obtain a solution that reaches a height of order one from initial data of small order. This is a beautiful example of a violation of the maximum principle for convection-diffusion problems of higher order. An open theoretical problem is to prove that with the nonlinear diffusion in (1.1), the undercompressive waves have such singular limits as $b \to 0$.

Numerical simulations of Münch [BMS99, M99] show that the undercompressive wave is the limit of a cascade of bifurcations that occur as the shock speed varies in (2.11). In particular, for small values of D, the phase portrait of the ODE at the critical speed s_* at which the undercompressive wave occurs has unusual structure. The unstable manifold from T is part of the topological boundary of the (two-dimensional) unstable manifold from M, which wraps around the unstable manifold from T in a spiral with an infinite number of turns. The result is that at the critical speed s_*, not only does the unstable manifold from T connect to B but the unstable manifold

from M intersects the stable manifold of B an infinite number of times; there are an infinite number of compressive waves connecting M to B with the undercompressive wave from T to B as their limit. Part of this structure is reminiscent of Silnikov's example [GH86] and we expect that machinery to be useful in studying this problem.

Finally we note that stability of traveling waves yields another interesting set of problems. Numerical simulations show that the undercompressive traveling waves, as solutions of (1.1), are stable with respect to perturbations. However, when there are multiple compressive waves at the same speed and with identical far field states, then some are stable and some are unstable.

A physically relevant problem is to gain more theoretical insight into the stability of traveling waves as plane wave solutions of the two-dimensional PDE

$$\partial_t h + \partial_x(h^2 - h^3) = D\nabla \cdot (h^3 \nabla h) - \nabla \cdot (h^3 \nabla \Delta h). \tag{6.3}$$

This is an important problem for understanding fingering patterns in driven film flow. Numerical simulations (with small D) show that compressive waves are typically unstable to transverse perturbations [THSJ89, BB97, KT97] while undercompressive waves are stable to transverse perturbations [BMFC98, KT98]. These stability differences are reflected in recent and ongoing experiments.

Recent progress has been made in understanding stability of undercompressive waves in systems of conservation laws [GZ98, LZ95]. These techniques will be used to explore stability to one- and two-dimensional perturbations from a more theoretical point of view in the near future [BMSZ99].

Acknowledgments. We thank Andrew Bernoff, Xiao-Biao Lin, Andreas Münch, Steve Schecter, and Kevin Zumbrun for useful conversations.

REFERENCES

[AK91] R. Abeyaratne and J. K. Knowles, *Kinetic relations and the propagation of phase boundaries in solids*, Arch. Rational Mech. Anal., 114 (1991), pp. 119–154.

[BB97] A. L. Bertozzi and M. P. Brenner, *Linear stability and transient growth in driven contact lines*, Phys. Fluids, 9 (1997), pp. 530–539.

[Ber98] A. L. Bertozzi, *The mathematics of moving contact lines in thin liquid films*, Notices Amer. Math. Soc., 45 (1998), pp. 689–697.

[BHP96] E. Beretta, J. Hulshof, and L. A. Peletier, *On an ODE from forced coating flow*, J. Differential Equations, 130 (1996), pp. 247–265.

[BMFC98] A. L. Bertozzi, A. Münch, X. Fanton, and A. M. Cazabat, *Contact line stability and 'undercompressive shocks' in driven thin film flow*, Phys. Rev. Lett., 81 (1998), pp. 5169–5172.

[BMS99] A. L. Bertozzi, A. Münch, and M. Shearer, *Undercompressive shocks in thin film flow*, Phys. D, 134 (1999), pp. 431–464.

[BMSZ99] A. L. Bertozzi, A. Münch, M. Shearer, and K. Zumbrun, *Stability of compressive and undercompressive thin film traveling waves*, European J. Appl. Math., submitted.

[CHTC90] A. M. Cazabat, F. Heslot, S. M. Troian, and P. Carles, *Finger instability of thin spreading films driven by temperature gradients*, Nature, 346 (1990), pp. 824–826.

[dB92] J. R. de Bruyn, *Growth of fingers at a driven three–phase contact line*, Phys. Rev. A, 46 (1992), pp. R4500–R4503.

[Fan98] X. Fanton, *Etalement et instabilités de films de mouillage en presence de gradients de tension superficielle*, Ph.D. thesis, Université Paris 6 Pierre et Marie Curie, Paris, 1998.

[Fre97] H. Freistühler, *Contributions to the mathematical theory of magnetohydrodynamic shock waves*, in Nonlinear Evolutionary Partial Differential Equations (Beijing, 1993), AMS, Providence, RI, 1997, pp. 175–187.

[GH86] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, 2nd ed., Springer–Verlag, New York, 1986.

[Gre78] H. P. Greenspan, *On the motion of a small viscous droplet that wets a surface*, J. Fluid Mech., 84 (1978), pp. 125–143.

[Gri89] M. Grinfeld, *Nonisothermal dynamic phase transitions*, Quart. Appl. Math., 47 (1989), pp. 71–84.

[GZ98] R. A. Gardner and K. Zumbrun, *The gap lemma and geometric criteria for stability of viscous shock profiles*, Comm. Pure Appl. Math., 51 (1998), pp. 789–847.

[HL97] B. T. Hayes and P. G. LeFloch, *Nonclassical shock waves: Scalar conservation laws*, Arch. Rational Mech. Anal., 139 (1997), pp. 1–56.

[HS98] B. Hayes and M. Shearer, *Undercompressive shocks for scalar conservation laws with non-convex fluxes*, Proc. Royal Soc. Edinburgh Sect. A, 129 (1999), pp. 733–754.

[Hup82] H. Huppert, *Flow and instability of a viscous current down a slope*, Nature, 300 (1982), pp. 427–429.

[IMP90] E. L. Isaacson, D. Marchesin, and B. J. Plohr, *Transitional waves for conservation laws*, SIAM J. Math. Anal., 21 (1990), pp. 837–866.

[IMPT92] E. L. Isaacson, D. Marchesin, B. J. Plohr, and J. B. Temple, *Multiphase flow models with singular Riemann problems*, Mat. Apl. Comput., 11 (1992), pp. 147–166.

[Jam80] R. D. James, *The propagation of phase boundaries in elastic bars*, Arch. Rational Mech. Anal., 73 (1980), pp. 125–158.

[JdB92] J. M. Jerrett and J. R. de Bruyn, *Finger instability of a gravitationally driven contact line*, Phys. Fluids A, 4 (1992), pp. 234–242.

[JMS95] D. Jacobs, B. McKinney, and M. Shearer, *Travelling wave solutions of the modified Korteweg-deVries-Burgers equation*, J. Differential Equations, 116 (1995), pp. 448–467.

[JSMB98] M. F. G. Johnson, R. A. Schluter, M. J. Miksis, and S. G. Bankoff, *Experimental study of rivulet formation on an inclined plate by fluorescent imaging,* J. Fluid Mech., 394 (1999), pp. 339–354.

[KH75] N. Kopell and L. N. Howard, *Bifurcations and trajectories joining critical points*, Adv. Math., 18 (1975), pp. 306–358.

[KT97] D. E. Kataoka and S. M. Troian, *A theoretical study of instabilities at the advancing front of thermally driven coating films*, J. Coll. Int. Sci., 192 (1997), pp. 350–362.

[KT98] D. E. Kataoka and S. M. Troian, *Stabilizing the advancing front of thermally driven climbing films*, J. Coll. Int. Sci., 203 (1998), pp. 335–344.

[LL71] V. Ludviksson and E. N. Lightfoot, *The dynamics of thin liquid films in the presence of surface-tension gradients*, Am. Inst. Chem. Engrg. J., 17 (1971), pp. 1166–1173.

[LZ95] T.-P. Liu and K. Zumbrun, *Nonlinear stability of general undercompressive shock waves*, Comm. Math. Phys., 174 (1995), pp. 319–345.

[M99] A. Münch, *Shock transitions in Marangoni-gravity driven thin film flow*, Nonlinearity, 13 (2000), pp. 731–746.

[MB99] A. Münch and A. L. Bertozzi, *Rarefaction-undercompressive fronts in driven films*, Phys. Fluids, 11 (1999), pp. 2812–2814.

[Mic88] D. Michelson, *Strong viscous shocks for systems of conservation laws with a high order of dissipation*, J. Differential Equations, 71 (1988), pp. 246–254.

[Ren96] M. Renardy, *A singularly perturbed problem related to surfactant spreading on thin films*, Nonlinear Anal., 27 (1996), pp. 287–296.

[She86] M. Shearer, *Nonuniqueness of admissible solutions of Riemann initial value problems for a system of conservation laws of mixed type*, Arch. Rational Mech. Anal., 93 (1986), pp. 45–59.

[Sle83] M. Slemrod, *Admissibility criteria for propagating phase boundaries in a van der Waals fluid*, Arch. Rational Mech. Anal., 81 (1983), pp. 301–315.

[Smo94] J. Smoller, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, New York, 2nd ed., 1994.

[SSMPL87] M. Shearer, D. G. Schaeffer, D. Marchesin, and P. Paes-Leme, *Solution of the Riemann problem for a prototype* 2×2 *system of non-strictly hyperbolic conservation laws*, Arch. Rational Mech. Anal., 97 (1987), pp. 299–320.

[Tay96] M. E. Taylor, *Partial Differential Equations* III*, Nonlinear Equations*, Appl. Math. Sci. 117, Springer–Verlag, New York, 1996.

[THSJ89] S. M. Troian, E. Herbolzheimer, S. A. Safran, and J. F. Joanny, *Fingering instabilities of driven spreading films*, Europhys. Lett., 10 (1989), pp. 25–30.

[Tru87] L. M. Truskinovsky, *Dynamics of non-equilibrium phase boundaries in a heat conducting non-linearly elastic medium*, PMM U.S.S.R., 51 (1987), pp. 777–784.

[VIC98] I. Veretennikov, A. Indeikina, and H.-C. Chang, *Front dynamics and fingering of a driven contact line*, J. Fluid Mech., 373 (1998), pp. 81–110.

SIAM J. MATH. ANAL.
Vol. 32, No. 1, pp. 214–233

ON SOLVABILITY OF THE NONLINEAR WAVE RESISTANCE PROBLEM FOR A SURFACE-PIERCING SYMMETRIC CYLINDER*

CARLO D. PAGANI† AND DARIO PIEROTTI†

Abstract. We solve the wave-resistance problem for a "slender" cylinder semisubmerged in a heavy fluid and moving at uniform, supercritical speed in the direction orthogonal to its generators. We prove that the free boundary and the cylinder profile form a single smooth streamline; moreover, the free boundary is monotone increasing downstream and lies under the level of calm water.

Key words. free boundary, nonlinear boundary condition, hodograph transformation

AMS subject classifications. 35J65, 35R35, 76B10

PII. S0036141098348140

1. Introduction. Consider a rigid body moving at a uniform speed c on the free surface of a heavy fluid. The unperturbed fluid, which is at rest, is assumed to have a finite constant depth H; compressibility and viscosity are neglected, as is surface tension; moreover, the fluid motion is assumed to be irrotational (all these assumptions are common in the theory of surface gravity waves). The *wave resistance problem* consists of the determination of the steady flow generated by this motion. Notice that we neglect the effect that the perturbed fluid produces on the motion of the body (sea-keeping problem). Besides the general hypotheses stated above, we make three main assumptions:

(a) *Two-dimensional geometry.* The body is an infinitely long, horizontal cylinder moving in the direction orthogonal to its generators and producing two-dimensional disturbances on the fluid which can be completely described in a vertical plane containing the direction of the motion (see Figure 1).

(b) *Supercritical velocity.* We assume

$$c > \sqrt{gH}, \tag{1.1}$$

where g is the acceleration of gravity. As we will see, this condition allows us to consider a flow which is unperturbed at infinity in *both* directions; actually, in the case of subcritical velocity, nontrivial oscillations at backward infinity cannot be excluded (also at the level of the linearized problem; see [1]).

(c) *Slenderness of the body.* We assume that the piercing part of the cylinder is small compared to its length. To be precise, let us choose a reference frame connected with the cylinder and such that the xy-plane is orthogonal to the horizontal generators of the cylinder; the x-axis is directed as the unperturbed flow, the undisturbed free surface is at $y = 0$, and the bottom of the region occupied by the fluid is at $y = -H$. We formulate our assumption by saying that the boundary of the cross section of the "hull" is described by the equation

$$y = \epsilon f(x),$$

*Received by the editors November 25, 1998; accepted for publication (in revised form) November 4, 1999; published electronically June 22, 2000.
http://www.siam.org/journals/sima/32-1/34814.html
†Dipartimento di Matematica del Politecnico, Piazza Leonardo da Vinci 32, 20133 Milano, Italy (darpie@mate.polimi.it, carpag@mate.polimi.it).

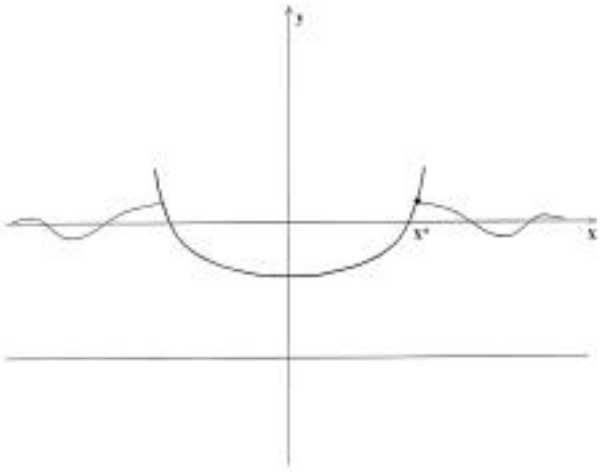


FIG. 1. *Vertical section of the system fluid-cylinder.*

where $\epsilon > 0$ is a small parameter and f is a $\mathcal{C}^1$ symmetric function defined in some interval $J \in \mathbf{R}$ and such that, for some $x_0 > 0$, with $\pm x_0 \in J$,

$$f(x) < 0 \qquad \text{for} \quad x \in (-x_0, x_0),$$

$$f(\pm x_0) = 0,$$

$$f(x) > 0 \qquad \text{for} \quad x \in J \backslash [-x_0, x_0].$$

The choice of a symmetric profile for the cylinder section allows us to simplify the problem by looking for solutions with a definite symmetry (with respect to x) and by reducing the number of unknown parameters (see below). The general, nonsymmetric case will be treated in a forthcoming paper. Finally, it remains to specify the properties of the free surface of the fluid: we assume that it is described by the cartesian curve $y = h(x)$, where h is an unknown symmetric smooth function defined in $\mathbf{R}\backslash[-x^*, x^*]$, for some positive $x^* \in J$. The numbers $\pm x^*$ are the abscissae of the points where the free surface meets the hull, so that $h(\pm x^*) = \epsilon f(\pm x^*)$; clearly, we will also require $h(x) < \epsilon f(x)$ for every $x \in J$ with $|x| > x^*$. Note that the value x^* is unknown and its determination is part of the problem. We set

$$h^*(x) = \begin{cases} h(x) & \text{for } |x| > x^*, \\ \epsilon f(x) & \text{for } |x| \le x^*. \end{cases}$$

Then

$$S^* = \{(x, y) \in \mathbf{R}^2 \, : -H < y < h^*(x)\} \tag{1.2}$$

will denote the region filled with the fluid. We assume (as usual) that the curve $y = h^*(x)$ is a streamline; also, the bottom $y = -H$ is assumed to be a streamline.

To formulate the various equations of the problem, it is convenient to introduce the complex variable $z = x + iy$ and the complex velocity function $\omega(z) = u(x, y) - iv(x, y)$, holomorphic in S^*, with u and v components of the velocity vector. We can now state our problem in the following form: find a real number $x^* > 0$, a real symmetric

function $h \in \mathcal{C}^1(\mathbf{R})$, and a complex function $\omega = u - iv$ holomorphic in S^*, with $\omega(-x+iy) = \bar{\omega}(x+iy)$, such that the following boundary conditions hold:

$$\frac{1}{2}|\omega(x,h(x))|^2 + gh(x) = \text{constant}, \qquad |x| > x^*, \tag{1.3}$$

$$v(x,h(x)) = h'(x)u(x,h(x)), \qquad |x| > x^*, \tag{1.4}$$

$$v(x,\epsilon f(x)) = \epsilon f'(x)u(x,\epsilon f(x)), \qquad |x| < x^*, \tag{1.5}$$

$$v(x,-H) = 0, \qquad x \in \mathbf{R}, \tag{1.6}$$

$$\lim_{|x|\to\infty} \omega(z) = c; \tag{1.7}$$

$$\lim_{|x|\to\infty} h(x) = 0. \tag{1.8}$$

Equations (1.4), (1.5) indicate that the free surface and the wetted hull are arcs of a streamline; (1.6) expresses the same property for the bottom, while (1.3) is the Bernoulli condition on the free surface. Finally, we have the *continuity condition*

$$h(x^*) = \epsilon f(x^*), \tag{1.9}$$

together with the inequality

$$h(x) < \epsilon f(x) \qquad \text{for} \quad x \in J\backslash[-x^*, x^*]. \tag{1.10}$$

Notice that, by using (1.7) and (1.8) in the Bernoulli condition (1.3), one obtains for the constant at the right-hand side the value $\frac{1}{2}c^2$.

The problem outlined above appears in the literature in its *linearized version* (called Neumann–Kelvin problem). This is an artificial problem, obtained by assuming that the profile of the free boundary is a small perturbation of the flat surface, regardless of the size or the shape of the moving body. In this way, one is faced with a linear problem (for the velocity potential or the stream function [1]) in a *fixed* domain; however, this problem is inconsistent, in the sense that it admits an infinite set of solutions depending upon two real parameters (one parameter in the symmetric case). Several different supplementary conditions were proposed [1], [2], [3]; as was noted in [2], all of them give mathematically well-posed statements, but it is not clear which of them better meets the hydrodynamics of the phenomenon.

The aim of this paper is to show the existence of an exact solution of the nonlinear problem for small values of the parameter ϵ, which in the limit case of a beam, i.e., for $\epsilon \to 0$, reduces to the free parallel flow $\omega(z) = c$, $h(x) = 0$.

2. The hodograph transformation. It is convenient to reformulate the problem (as we did in [4]) by using as new independent variables the velocity potential $\varphi = \varphi(x,y)$ and the stream function $\psi = \psi(x,y)$. Let w be the complex potential

$$w = \varphi + i\psi, \qquad w'(z) = \omega(z). \tag{2.1}$$

Then,

$$u = \varphi_x = \psi_y, \qquad v = \varphi_y = -\psi_x. \tag{2.2}$$

The lines $\varphi =$ constant are the equipotential lines, while $\psi =$ constant are the streamlines. The region S^* being simply connected, the potential w is determined by the complex velocity up to an additive constant. Let us now fix the complex constant: the real part is fixed by requiring that $\varphi(0, y) = 0$ (so that φ is antisymmetric with respect to x); the imaginary part is fixed by requiring that the streamline $y = h^*(x)$ is represented by the equation $\psi = 0$. As a consequence, the bottom $y = -H$ will be described by the equation $\psi = -cH$, for we have

$$\psi(x, -H) = -\int_{-H}^{h^*(x)} \psi_y(x, t)dt = -\int_{-H}^{h^*(x)} u(x, t)dt.$$

The last integral represents the total flux crossing any vertical section of S^* and is independent of x; at infinity, where the flow is unperturbed, it becomes $-\int_{-H}^{0} c dt = -cH$.

We recall now that we are looking for a solution which is a small perturbation of the free parallel flow; hence, it is reasonable to assume that $u(x, y) > 0$ uniformly in S^*, and therefore $\omega(z) \neq 0$ in S^*. Moreover, by the first equation of (2.2), the maps $x \mapsto \varphi(x, y)$ and $y \mapsto \psi(x, y)$ are strictly increasing. It follows that there is a conformal map, called the hodograph

$$z \mapsto w(z) \tag{2.3}$$

which maps the domain S^* of the physical plane onto a strip A_H in the hodograph plane (φ, ψ) given by

$$A_H \equiv \{(\varphi, \psi) \in \mathbf{R}^2 : \quad -cH < \psi < 0\}. \tag{2.4}$$

We may also assume that w is one-to-one, so that the inverse map

$$w \mapsto z(w)$$

is well defined on A_H and satisfies

$$\frac{dz}{dw} = \frac{1}{\omega(z)} \equiv \Omega(w). \tag{2.5}$$

As already discussed in [4], in the hodograph plane the flow is better described by the function Ω. By writing $\Omega = U - iV$, the above relation takes the form

$$U = \frac{\partial x}{\partial \varphi} = \frac{\partial y}{\partial \psi}, \qquad V = \frac{\partial x}{\partial \psi} = -\frac{\partial y}{\partial \varphi}. \tag{2.6}$$

By noticing that

$$U = \frac{u}{u^2 + v^2}, \qquad V = -\frac{v}{u^2 + v^2} \tag{2.7}$$

and by the above symmetry assumptions, we easily verify the relations

$$\Omega(-\varphi, \psi) = \bar{\Omega}(\varphi, \psi),$$

and

$$\Omega \to 1/c$$

for $|\varphi| \to \infty$. Then, we can write explicitly

$$x(\varphi,\psi) = \int_0^{\varphi} U(s,\psi)ds, \qquad y(\varphi,\psi) = \frac{1}{c}\psi + \int_{\varphi}^{+\infty} V(s,\psi)ds, \tag{2.8}$$

assuming that the last integral is convergent.

We now specify the boundary conditions that the holomorphic function Ω has to satisfy on ∂A_H. We note first that two points are relevant on the upper boundary of the strip A_H: the images by the hodograph map of the points $P_\pm = (\pm x^*, \epsilon f(x^*))$ where the free boundary meets the hull; let us write $\varphi = \pm\varphi^*$ for the equipotential lines passing through these points. The value of φ^* is unknown and its determination is part of our problem. The arc of the curve between P_+ and P_-, i.e., the wetted part of the hull, is mapped onto the beam

$$I = \{(\varphi,\psi) : \quad \psi = 0, \quad |\varphi| < \varphi^*\}, \tag{2.9}$$

and the free surface onto the half-lines

$$F = \{(\varphi,\psi) : \quad \psi = 0, \quad |\varphi| > \varphi^*\}. \tag{2.10}$$

Furthermore, the image of the bottom of the fluid is the line

$$B = \{(\varphi,\psi) : \quad \psi = -cH, \quad \varphi \in \mathbf{R}\}. \tag{2.11}$$

Now, we observe that the kinematic free surface condition (1.4) is already taken into account by requiring that the free surface is part of the streamline $\psi = 0$, while the Bernoulli condition (1.3), by standard computations, takes the form

$$\frac{1}{2}|\Omega|^{-4}\frac{\partial|\Omega|^2}{\partial\varphi} + gV = 0 \qquad \text{on } F. \tag{2.12}$$

The condition (1.4) becomes

$$V(\varphi,0) + \epsilon f'(x(\varphi,0))U(\varphi,0) = 0 \qquad \text{for } |\varphi| < \varphi^*, \tag{2.13}$$

while the condition on the bottom gives

$$V = 0 \qquad \text{on } B. \tag{2.14}$$

Moreover, we require the asymptotic condition

$$\lim_{|\varphi|\to\infty} \Omega = \frac{1}{c}. \tag{2.15}$$

Finally, by taking into account (2.8), the continuity condition (1.9) has the form

$$\int_{\varphi^*}^{+\infty} V(s,0)ds = \epsilon f\left(\int_0^{\varphi^*} U(s,0)ds\right). \tag{2.16}$$

Equations (2.12)–(2.16) formulate the problem in the hodograph plane.

We point out that the free surface profile, $h(x)$, disappeared among the unknowns; it will be recovered at the end, once the hodograph map is known, as the image of the level line $\psi = 0$. We also notice that (2.12)–(2.15) is a boundary value problem for a

function Ω holomorphic in a *fixed* domain, namely the strip A_H; however, the position of the points $(\pm\varphi^*, 0)$ separating the two different boundary conditions (2.12), (2.13) is unknown.

To solve the problem, we choose the following strategy: first we regard φ^* as a known parameter and find the holomorphic function Ω satisfying (2.12)–(2.15) by means of the implicit function theorem in Banach spaces; then, we put the above solution (which will depend on φ^*) in the continuity condition (2.16) and obtain an equation for the unknown φ^*. The necessity of this two-step procedure is due to the fact that we do not know the limit positions for $\epsilon \to 0$ of the abscissae of the points $P_\pm$, so that we cannot linearize the whole problem for Ω, φ^* around the solution at $\epsilon = 0$.

Assume then that φ^* is known; a rescaling of the independent variables and a further change of the unknowns will prove convenient in the following. By setting

$$\rho = \frac{\varphi}{\varphi^*}, \qquad \sigma = \frac{\psi}{\varphi^*}, \qquad \zeta = \rho + i\sigma, \tag{2.17}$$

the strip A_H becomes

$$A^* \equiv \left\{ (\rho, \sigma) \in \mathbf{R}^2 \, : \quad -\frac{cH}{\varphi^*} < \sigma < 0 \right\}. \tag{2.18}$$

In particular, the beam (2.9) maps onto the interval $(-1, 1)$ of the ρ-axis.

We now observe that, for $\epsilon = 0$, (2.12)–(2.15) admit the constant solution $\Omega = 1/c$. Then, we define the new unknown $\chi = \xi - i\eta$ (as functions of the rescaled variables (2.17)) by subtracting this solution from Ω and dividing by ϵ; namely, we set

$$U(\varphi, \psi) = \frac{1}{c}\Big(1 + \epsilon\xi(\rho, \sigma)\Big), \qquad V(\varphi, \psi) = \frac{\epsilon}{c}\eta(\rho, \sigma). \tag{2.19}$$

We want to write the nonlinear boundary conditions (2.12), (2.13) as formal operator equations in the new variables. We first note that the relations (2.8) take the form

$$x(\varphi, \psi) = \frac{\varphi^*}{c}\int_0^\rho (1 + \epsilon\xi(s, \sigma))ds, \qquad y(\varphi, \psi) = \frac{\varphi^*}{c}\left\{\sigma + \epsilon\int_\rho^{+\infty} \eta(s, \sigma)ds\right\}, \tag{2.20}$$

and that we can define on $(-1, 1)$ the function

$$G(\rho) = f'(x(\varphi, 0)) = f'\left(\frac{\varphi^*}{c}\int_0^\rho (1 + \epsilon\xi(s, 0))ds\right). \tag{2.21}$$

We now set

$$B^I(\chi, \epsilon) = \{\eta + G(\cdot)(1 + \epsilon\xi)\}|_{|\rho|<1,\, \sigma=0}, \tag{2.22}$$

$$B^F(\chi, \epsilon) = \left\{\frac{|1 + \epsilon\chi|^{-4}}{2\epsilon}\frac{\partial}{\partial\rho}|1 + \epsilon\chi|^2 + \frac{g\varphi^*}{c^3}\eta\right\}\Big|_{|\rho|>1,\, \sigma=0}, \tag{2.23}$$

and

$$\mathbf{B}(\chi, \epsilon) = (B^I(\chi, \epsilon), B^F(\chi, \epsilon)). \tag{2.24}$$

Then, it is easily verified that the equation

$$\mathbf{B}(\chi, \epsilon) = 0 \tag{2.25}$$

is equivalent to the conditions (2.12)–(2.13). Moreover, the function χ must be holomorphic in A^*, vanishing for $|\rho| \to \infty$ and satisfying the *linear* condition

$$\eta(\rho, -cH/\varphi^*) = 0.$$

In the next section, we will formulate (2.25) as an operator equation between suitable Banach spaces which takes into account all of the above conditions.

3. The functional setting of the problem. We shall discuss first (section 3.1) a linear problem obtained from (2.25) by letting $\epsilon \to 0$ (formally). The results obtained will suggest the correct functional setting of the nonlinear problem (section 3.2). We recall that we are assuming here that φ^* is a known parameter.

3.1. The linearized problem. As we have already remarked, when $\epsilon = 0$, (2.12)–(2.15) admit the trivial solution $\Omega = 1/c$. Now we assume that Ω can be expanded in powers of ϵ and, according to (2.19), we set

$$\chi(\rho, \sigma) = \tilde{\chi}(\rho, \sigma) + \mathcal{O}(\epsilon); \tag{3.1}$$

by inserting (3.1) into (2.22), (2.23) and by taking the limit $\epsilon \to 0$, we get a problem satisfied by the holomorphic function $\tilde{\chi} = \tilde{\xi} - i\tilde{\eta}$ in the fixed domain A^* (see [5] for the details):

$$\tilde{\xi}_\rho + \frac{\varphi^* g}{c^3}\tilde{\eta} = 0 \qquad \text{for} \quad \sigma = 0, \ \ |\rho| > 1, \tag{3.2}$$

$$\tilde{\eta}(\rho, 0) = -f'\left(\frac{\varphi^*}{c}\rho\right) \qquad \text{for} \quad |\rho| < 1, \tag{3.3}$$

$$\tilde{\eta} = 0 \qquad \text{for} \quad \sigma = -\frac{cH}{\varphi^*}, \ \ \rho \in \mathbf{R}, \tag{3.4}$$

$$\lim_{|\rho| \to \infty} \tilde{\chi} = 0. \tag{3.5}$$

By substituting, in (3.2), $\tilde{\xi}_\rho$ with $-\tilde{\eta}_\sigma$, we obtain a boundary value problem for the harmonic function $\tilde{\eta}$ (the harmonic conjugate $\tilde{\xi}$ is then determined by the requirement of vanishing at infinity).

Problem L. Find $\tilde{\eta}$ harmonic in A^* such that

$$\tilde{\eta}_\sigma - \nu^* \tilde{\eta} = 0 \qquad \text{for} \quad \sigma = 0, \ \ |\rho| > 1,$$

$$\tilde{\eta}(\rho, 0) = -f'\left(\frac{\varphi^*}{c}\rho\right) \qquad \text{for} \quad |\rho| < 1,$$

$$\tilde{\eta} = 0 \qquad \text{for} \quad \sigma = -\frac{cH}{\varphi^*}, \ \ \rho \in \mathbf{R},$$

$$\lim_{|\rho| \to \infty} \tilde{\eta} = 0,$$

where we set

$$\nu^* = \frac{\varphi^* g}{c^3}.$$

Problem L coincides with the problem (1.1)–(1.4) in [5], which was obtained by linearizing (formally) the original nonlinear problem in the physical plane.

By the results of [5, section 3], it turns out that, for every symmetric and smooth enough function f, there exists a unique solution $\tilde{\eta} \in H^1(A^*)$ of the problem L and furthermore, a unique harmonic conjugate $\tilde{\xi}$ also belonging to $H^1(A^*)$; in addition, these functions are continuous in $\bar{A}^*$ and rapidly decreasing at infinity.

Our strategy is to find an appropriate functional setting for the equation $\mathbf{B}(\chi, \epsilon) = 0$ in order to apply the implicit function theorem. To do so, we need further properties of the solutions of the above linear problem. In fact, we must specify a Banach space norm on the space of these solutions in such a way that the boundary operators B^F and B^I are continuously differentiable; as we will show below, this requirement can be satisfied in the Sobolev space $W_p^2(A^*)$ for some $p > 1$. Moreover, we also have to solve the problem L with nonzero boundary data on $|\rho| > 1, \sigma = 0$, in order to prove the invertibility of the Fréchet derivative of the operator $\mathbf{B}$ (see below).

Then, we consider the following boundary value problem:

$$\Delta\tilde{\eta} = 0 \qquad \text{in} \quad A^*, \tag{3.6}$$

$$\tilde{\eta} = k \qquad \text{for} \quad \sigma = 0, \ \ |\rho| < 1, \tag{3.7}$$

$$\tilde{\eta}_\sigma - \nu^* \tilde{\eta} = l \qquad \text{for} \quad \sigma = 0, \ \ |\rho| > 1, \tag{3.8}$$

$$\tilde{\eta} = 0 \qquad \text{for} \quad \sigma = -\frac{cH}{\varphi^*}, \ \ \rho \in \mathbf{R}. \tag{3.9}$$

The key result of this section is the following theorem.

THEOREM 3.1. *Assume that $k \in W_p^{2-\frac{1}{p}}(-1,1)$ and $l \in W_p^{1-\frac{1}{p}}(\mathbf{R}\backslash[-1,1])$ with $p \in (1, 4/3)$. Then, there exists a unique solution $\tilde{\eta} \in W_p^2(A^*)$ of the problem* (3.6)–(3.9).

Proof. We first note that, by Sobolev imbeddings, we have

$$W_p^{2-\frac{1}{p}}(-1,1) \subset H^{1/2}(-1,1)$$

and

$$W_p^{1-\frac{1}{p}}(\mathbf{R}\backslash[-1,1]) \subset H^{-1/2}(\mathbf{R}\backslash[-1,1]).$$

Then, by the same arguments as in Theorem 3.1 of [5], there exists a solution $\tilde{\eta} \in H^1$ of (3.6)–(3.9); if we further assume that l has compact support, we also have that $\tilde{\eta}(\rho, \sigma)$ is smooth for $\rho \in \mathbf{R}\backslash\operatorname{supp} l$ up to the boundary $\sigma = 0$ and it is rapidly decreasing for $|\rho| \to \infty$ (see the proof of Proposition 3.2 of [5]). Let us now take M such that $\operatorname{supp} l \subset (-M, M)$ and consider the regularity of $\tilde{\eta}$ in the rectangle $(-M, M)\times(-cH/\varphi^*, 0)$. In order to apply the regularity theory in polygonal domains, we observe that at the upper side of the rectangle the solution $\tilde{\eta}$ has a Dirichlet datum in $W_p^{2-\frac{1}{p}}$ for $|\rho| < 1$ and a Neumann datum $\eta_\sigma = \nu^*\eta + l \in W_p^{1-\frac{1}{p}}$ for $1 < |\rho| < M$; the second property follows by the Sobolev embedding $H^{1/2}(\mathcal{I}) \subset W_p^{1-\frac{1}{p}}(\mathcal{I})$, which

holds for every $p < 2$ and for every bounded interval $\mathcal{I}$. Then, we can apply Theorem 4.4.3.7 of [6] and conclude that $\tilde{\eta} \in W^2_p((-M, M)\times(-cH/\varphi, 0))$ for every $p \in (1, 4/3)$. By the above discussion, we also have $\tilde{\eta} \in W^2_p(A^*)$. Finally, for generic $l \in W_p^{1-\frac{1}{p}}$, we consider a sequence l_m of functions with compact supports, such that $l_m \to l$ in $W_p^{1-\frac{1}{p}}$ for $m \to \infty$. Then, by known a priori estimates in an infinite strip [6, section 4.2], the corresponding solutions η_m converge in $W^2_p(A^*)$ to the solution $\tilde{\eta}$ of (3.6)–(3.9). □

REMARK 3.2. (i) *We recall that for every $p > 1$ the inclusion $W^2_p(A^*) \subset \mathcal{C}^{0,\alpha}(\bar{A^*})$ holds with $\alpha = 2-2/p$; moreover, for $p > 1$ the space $W^2_p(A^*)$ is an algebra, a crucial property for the functional setting of the nonlinear problem (see below).*

(ii) *Roughly speaking, the reason for the limitation $p < 4/3$ is the following: in the neighborhood of the points $(1, 0)$ and $(-1, 0)$, the solution of the above problem could have a "singular" part proportional to $S(r, \theta) = r^{1/2}\sin(\theta/2)$, where r, θ equal the polar coordinates around the points (see* [5, section 3]*). Then, it is not difficult to check that the function S does not belong to W^2_p in a neighborhood of the origin for $p \geq 4/3$.*

(iii) *Clearly, more regularity of the datum l implies moreregularity for the trace of the solution on $\sigma = 0$, $|\rho| > 1$.*

By assuming further properties on the data, namely that k and l are odd functions, and l is rapidly decreasing at infinity, we can achieve a useful result for the discussion of the nonlinear problem. More precisely, let $\lambda_1 > 0$ be the first positive solution of

$$\tan(\lambda H) = \frac{\lambda c^2}{g}$$

(see [5, Proposition 3.2]). Then, we get the following corollary.

COROLLARY 3.3. *Assume that the functions k and l of Theorem* 3.1 *are odd, that $l \in \mathcal{C}^{0,\alpha}(\mathbf{R}\backslash(-\rho_0, \rho_0))$ for some $\rho_0 > 1$ (with $\alpha = 2 - 2/p$), and that*

$$\sup_{|\rho|\geq\rho_0} e^{\lambda^*|\rho|}|l(\rho)| < \infty,$$

with $\lambda^* = \varphi^*\lambda_1/c$. Then, there is a unique holomorphic function $\tilde{\chi} = \tilde{\xi} - i\tilde{\eta}$ belonging to $W^2_p(A^*)$, satisfying $\tilde{\chi}(-\rho, \sigma) = \overline{\tilde{\chi}(\rho, \sigma)}$ and the boundary conditions (3.7)–(3.9). *Furthermore, $\tilde{\chi}|_{\sigma=0} \in \mathcal{C}^{1,\alpha}(\mathbf{R}\backslash(-\rho_0, \rho_0))$ and the following bounds hold:*

$$\sup_{A^*} e^{\lambda^*|\rho|}|\tilde{\chi}(\rho, \sigma)| < \infty,$$

$$\sup_{|\rho|\geq\rho_0} e^{\lambda^*|\rho|}|\partial_\rho\tilde{\xi}(\rho, 0)| < \infty.$$

The proof is given in the appendix.

REMARK 3.4. *By recalling the relation $\tilde{\xi}_\rho = -\tilde{\eta}_\sigma$, which holds in A^*, we can rephrase the boundary condition* (3.8) *in the form*

$$\tilde{\xi}_\rho + \nu^*\tilde{\eta} = -l \tag{3.10}$$

for $\sigma = 0$, $|\rho| > 1$, where the above relation holds between elements of

$$W_p^{1-\frac{1}{p}}(\mathbf{R}\backslash[-1, 1]) \cap \mathcal{C}^{0,\alpha}(\mathbf{R}\backslash(-\rho_0, \rho_0)).$$

3.2. The solution of the nonlinear problem. Take $\rho_0 > 1$ and denote by $Q_0 \subset \bar{A}^*$ the closed region $[-H^*, 0] \times \mathbf{R}\backslash(-\rho_0, \rho_0)$. Let us define the following set:

$$X = \Big\{ \chi = \xi - i\eta \text{ holomorphic in } A^*, \quad \chi(-\rho, \sigma) = \overline{\chi(\rho, \sigma)},$$

$$\chi \in W_p^2(A^*) \cap \mathcal{C}^{1,\alpha}(Q_0), \quad \eta(\cdot, -\frac{cH}{\varphi^*}) = 0, \quad ||\chi||_* < \infty \Big\}, \tag{3.11}$$

where $1 < p < 4/3$ and

$$||\chi||_* = ||\chi||_{\mathcal{C}^{1,\alpha}(Q_{\rho_0})} + ||\chi||_{W_p^2(A^*)} + \sup_{A^*} e^{\lambda^*|\rho|}|\chi(\rho,\sigma)| + \sup_{Q_{\rho_0}} e^{\lambda^*|\rho|}|\partial_\rho \xi(\rho,\sigma)|.$$

It is not difficult to check that X is a Banach space of continuous functions vanishing at infinity. Let us denote by Y_0 the set of the (real) functions

$$l \in W_p^{1-\frac{1}{p}}(\mathbf{R}\backslash[-1,1]) \cap \mathcal{C}^{0,\alpha}(\mathbf{R}\backslash(-\rho_0, \rho_0)),$$

such that the norm

$$||l|| = ||l||_{W_p^{1-\frac{1}{p}}(\mathbf{R}\backslash[-1,1])} + ||l||_{\mathcal{C}^{0,\alpha}(\mathbf{R}\backslash(-\rho_0,\rho_0))} + \sup_{|\rho|\geq\rho_0} e^{\lambda^*|\rho|}|l(\rho)|,$$

is finite. We now define

$$Y = \{(k,l) \in W_p^{2-\frac{1}{p}}(-1,1) \times Y_0, \quad k(-\rho) = -k(\rho),\, l(-\rho) = -l(\rho)\}, \tag{3.12}$$

with the usual norm for tensor products of Banach spaces. Then, we can state the following.

THEOREM 3.5. *Let f be a symmetric $\mathcal{C}^2$ function defined in an interval $J \supset [-\frac{\varphi^*}{c}, \frac{\varphi^*}{c}]$ and suppose that the Nemitski operator associated to f'' is continuous from $W_p^{3-\frac{1}{p}}(-1,1)$ to $W_p^{2-\frac{1}{p}}(-1,1)$. Then, there exist $\epsilon_0 > 0$ and a bounded open set $\mathcal{U} \subset X$ containing the solution $\tilde{\chi}$ of problem (3.2)–(3.5), such that the operator*

$$\mathbf{B} : \mathcal{U} \times [0, \epsilon_0) \to Y$$

defined by (2.24) is continuously differentiable.

Proof. By recalling the expression (2.22) of B^I, we may choose ϵ_0 and $\mathcal{U}$ such that, if $(\chi, \epsilon) \in \mathcal{U} \times [0, \epsilon_0)$, the relation $\frac{\varphi^*}{c}\int_0^\rho [1 + \epsilon\xi(t,0)]dt \in J$ holds for every $\rho \in [-1,1]$. Then, by our assumptions on f'' and the continuity of the product between functions in $W_p^{2-\frac{1}{p}}(-1,1)$ (see [6, Theorem 1.4.4.2]), the derivative (G-differential) of B^I at $\chi^* = \xi^* - i\eta^*$ in the direction $\chi = \xi - i\eta$ exists and is equal to

$$d_G B^I(\chi^*, \epsilon)\chi = \eta(\rho, 0) - \epsilon f'\left(\frac{\varphi^*}{c}\int_0^\rho [1 + \epsilon\xi^*(t,0)]dt\right)\xi(\rho,0)$$

$$-\epsilon\frac{\varphi^*}{c} f''\left(\frac{\varphi^*}{c}\int_0^\rho [1 + \epsilon\xi^*(t,0)]dt\right)[1 + \epsilon\xi^*(\rho,0)]\int_0^\rho \xi \tag{3.13}$$

with $\rho \in (-1,1)$. Furthermore, the right-hand side of (3.13) defines a bounded linear operator $d_G B^I(\chi^*, \epsilon) : X \to W_p^{2-\frac{1}{p}}(-1,1)$ and one can easily check that the map

$(\chi^*, \epsilon) \mapsto d_G B^I(\chi^*, \epsilon)$ is continuous. Then B^I is Frechet differentiable with continuous derivative in $\mathcal{U} \times [0, \epsilon_0)$. The differentiability of B^I with respect to ϵ is readily verified.

Let us now consider the operator B^F given by (2.23) and take ϵ_0 small enough so that $\inf_{|\rho|\geq 1} |1 + \epsilon\chi| > 0$ for every $\chi \in \mathcal{U}$. Then, by a straighforward calculation we can write

$$B^F(\chi, \epsilon) = \Big\{ |1 + \epsilon\chi|^{-4} [\xi_\rho + \epsilon(\xi_\rho \xi + \eta_\rho \eta)] + \nu^* \eta \Big\}\Big|_{|\rho|>1, \sigma=0}. \tag{3.14}$$

By the above expression, by the continuity of the application $f, g \mapsto f \cdot g$ from $W_p^{2-\frac{1}{p}}(\mathbf{R}\backslash[-1,1]) \times W_p^{1-\frac{1}{p}}(\mathbf{R}\backslash[-1,1])$ into $W_p^{1-\frac{1}{p}}(\mathbf{R}\backslash[-1,1])$ (see [6, Theorem 1.4.4.2]), and by Corollary 3.3, we find that B^F is a well defined, continuous operator from X into Y. Moreover, the G-derivative at χ^* is given by

$$d_G B^F(\chi^*, \epsilon)\chi = \xi_\rho + \nu^* \eta + \mathcal{O}(\epsilon), \tag{3.15}$$

where $\mathcal{O}(\epsilon)$ represents a function depending on χ, χ^*, and their derivatives, whose norm, for $\epsilon \to 0$ (and χ, χ^* in a bounded set of X) is $\mathcal{O}(\epsilon)$. Hence, we obtain as before that

$$d_G B^F(\chi^*, \epsilon) \,:\, X \to Y_0$$

is a bounded linear operator and that the map $(\chi^*, \epsilon) \mapsto d_G B^F(\chi^*, \epsilon)$ is continuous in $\mathcal{U} \times [0, \epsilon_0)$. Finally, again from (3.14) we easily get the differentiability of B^F with respect to ϵ. □

REMARK 3.6. *A sufficient condition for the continuity of the Nemitski operator associated to f'' is that $f \in \mathcal{C}^{3,1}(J)$; for in this case we have $f'' \in W_\infty^2$, so that f, f', and f'' are all continuous from $L^p(J)$ to itself.*

By denoting with $\mathbf{B}' = \mathbf{B}'(\chi, \epsilon)$ the Frechet differential of $\mathbf{B}$ with respect to χ, we get from (3.13) *and* (3.15)

$$\mathbf{B}'(\chi^*, 0)\chi = \Big(\eta\Big|_{|\rho|<1, \sigma=0}, \{\xi_\rho + \nu^* \eta\}\Big|_{|\rho|>1, \sigma=0} \Big),$$

so that, by Theorem 3.1*, Corollary* 3.3 *and Remark* 3.4 *we get the following.*

COROLLARY 3.7. *For every $\chi \in \mathcal{U}$ the operator $\mathbf{B}'(\chi, 0)$ is invertible.*

Now, by applying the implicit function theorem, we obtain the following.

THEOREM 3.8. *Let f satisfy the assumptions of Theorem* 3.5*; then, there exists $\epsilon_0 > 0$ such that, for every $\epsilon \in [0, \epsilon_0)$, the equation $\mathbf{B}(\chi, \epsilon) = 0$ has a unique solution $\chi^\epsilon \in \mathcal{U}$. Moreover, the map $\epsilon \mapsto \chi^\epsilon$ is continuously differentiable.*

4. Proof of the main result. In order to prove the existence of a solution to our problem, we still have to solve the continuity condition (2.16). To do so, some properties of the solution obtained in the previous section are needed; in particular, we investigate its dependence on the parameter φ^*. First, we prove a result which is interesting in itself, since it implies that the free boundary $h(x)$ associated to the solution of our problem must be strictly increasing for x positive.

THEOREM 4.1. *For any fixed $\varphi^* > 0$ and for every small enough $\epsilon \geq 0$, the solution $\chi^\epsilon = \xi^\epsilon - i\eta^\epsilon$ given by Theorem* 3.8 *satisfies*

$$\eta^\epsilon(\rho, 0) < 0 \tag{4.1}$$

for every $\rho \in (0, +\infty)$.

Proof. Let us consider the function η^ϵ in the half strip $[0,+\infty)\times[-\frac{cH}{\varphi^*},0]$. We recall that η^ϵ is continuous, vanishes for $\rho=0$ by symmetry and for $\sigma=-\frac{cH}{\varphi^*}$; furthermore, by the condition $B^I(\chi^\epsilon,\epsilon)=0$, we have $\eta^\epsilon(\rho,0)<0$ for $0<\rho\le 1$ and ϵ small. Assume, for contradiction, that $\sup_{\rho\ge1}\eta^\epsilon(\rho,0)=M>0$; since η^ϵ goes to zero at infinity, it must be $M=\eta^\epsilon(\bar\rho,0)$ for some $\bar\rho>1$. Note that $\bar\rho=\bar\rho(\epsilon)>1$ uniformly for ϵ in a neighborhood of zero. Let us now consider the harmonic function

$$W_M=M\left(\frac{\varphi^*}{cH}\sigma+1\right)$$

and compare W_M with η^ϵ in the domain $A_R=[0,R]\times[-\frac{cH}{\varphi^*},0]$, $R>0$ (a similar procedure, based on comparison methods, has been used in [7]). By the above discussion and assumption, we easily get $W_M(\rho,0)\ge\eta^\epsilon(\rho,0)$ in $[0,R]$, $W_M(0,\sigma)\ge\eta^\epsilon(0,\sigma)=0$ in $[-\frac{cH}{\varphi^*},0]$, $W_M(\rho,-\frac{cH}{\varphi^*})=\eta^\epsilon(\rho,-\frac{cH}{\varphi^*})=0$ in $[0,R]$. Finally, since both $\eta^\epsilon(\rho,\sigma)$ and $\eta^\epsilon_\sigma(\rho,\sigma)$ are vanishing for $\rho\to+\infty$, we conclude that $W_M(R,\sigma)\ge\eta^\epsilon(R,\sigma)$ in $[-\frac{cH}{\varphi^*},0]$ for large enough R. By the maximum principle, $W_M\ge\eta^\epsilon$ in A_R and the equal sign holds at $(\bar\rho,0)$. Then, by the Hopf maximum principle,

$$\frac{\partial\eta^\epsilon}{\partial\sigma}(\bar\rho,0)>\frac{\partial W_M}{\partial\sigma}(\bar\rho,0)=\frac{\varphi^*}{cH}M. \tag{4.2}$$

By recalling (3.14), the boundary condition $B^F(\chi^\epsilon,\epsilon)=0$ at $(\bar\rho,0)$ takes the form

$$\xi^\epsilon_\rho(\bar\rho,0)[1+\epsilon\xi^\epsilon(\bar\rho,0)]+\epsilon\eta^\epsilon_\rho(\bar\rho,0)\eta^\epsilon(\bar\rho,0)+\nu^*\big|1+\epsilon\chi^\epsilon(\bar\rho,0)\big|^4\eta^\epsilon(\bar\rho,0)=0.$$

Now, since $\eta^\epsilon(\rho,0)$ is smooth (see Remark A.2) with a maximum M at $\bar\rho$, and by the relation $\xi_\rho=-\eta_\sigma$, we have

$$\frac{\partial\eta^\epsilon}{\partial\sigma}(\bar\rho,0)=\frac{g\varphi^*}{c^3}\frac{|1+\epsilon\chi^\epsilon(\bar\rho,0)|^4}{1+\epsilon\xi^\epsilon(\bar\rho,0)}M. \tag{4.3}$$

By inserting (4.3) in (4.2) we obtain the relation

$$\frac{c^2}{gH}\frac{1+\epsilon\xi^\epsilon(\bar\rho,0)}{|1+\epsilon\chi^\epsilon(\bar\rho,0)|^4}<1. \tag{4.4}$$

Then, since $c^2/gH>1$, we reach a contradiction for small enough ϵ. Finally, the case $M=0$ is excluded by (4.3) and again by the Hopf principle. □

REMARK 4.2. *From the above result and by recalling the relations* (2.20)*, it follows that the curve in the physical plane parametrized by $x(\varphi^*\rho,0)$ and $y(\varphi^*\rho,0)$ is negative and strictly increasing in $(0,+\infty)$. We remark that such a curve will represent the physical free boundary (for $|\rho|>1$) only if the continuity condition is satisfied.*

In the following, we will emphasize the dependence on φ^* of the operator $\mathbf{B}$ by writing $\mathbf{B}(\chi,\epsilon)=\mathbf{B}(\chi,\epsilon;\varphi^*)$; similarly, we set $\chi^\epsilon(\zeta)=\chi^\epsilon(\zeta;\varphi^*)$ for the solutions of $\mathbf{B}(\chi,\epsilon;\varphi^*)=0$. Notice that also the space X defined by (3.11) depends on φ^*; then, in order to compare solutions for different values of φ^*, it is convenient to discuss a suitable extension of $\mathbf{B}$.

Let us choose $\varphi^*_1>0$, φ^* in a neighborhood of φ^*_1 and define the strip A^*_1 and the space X_1 by setting $\varphi^*=\varphi^*_1$ in (2.18) and (3.11), respectively, but *dropping* the condition $\eta(\cdot,-cH/\varphi^*)=0$; we denote by the same symbol $\mathbf{B}(\cdot,\epsilon;\varphi^*)$ the operator defined in an open bounded set of X_1 by

$$\mathbf{B}(\chi,\epsilon;\varphi^*)=(B^I(\chi,\epsilon;\varphi^*),B^F(\chi,\epsilon;\varphi^*),\eta|_{\sigma=-cH/\varphi^*_1}), \tag{4.5}$$

with B^I, B^F, defined in (2.22)–(2.23). One verifies that the last component of $\mathbf{B}$ takes values in a space of smooth, rapidly decreasing functions on $\mathbf{R}$. We set

$$Y_* = \Big\{ l \in \mathcal{C}^{1,\alpha}(\mathbf{R}), \qquad \sup_{|\rho|\in\mathbf{R}} e^{\lambda^*|\rho|}(|l(\rho)| + |l'(\rho)|) < \infty \Big\}.$$

Note that the inclusion $Y_* \subset W_p^{2-\frac{1}{p}}(\mathbf{R})$ holds.

The following lemma extends some results of the previous section.

LEMMA 4.3. *Take $\varphi_1^* > 0$ such that $\varphi_1^*/c \in J$. Then, the map*

$$(\chi, \epsilon, \varphi^*) \mapsto \mathbf{B}(\chi, \epsilon; \varphi^*)$$

is continuously differentiable from $\mathcal{V} \times [0, \epsilon_0) \times \mathcal{J}$ into $Y \times Y_*$, where $\mathcal{V}$ is an open ball in X_1, $\epsilon_0 > 0$ and $\mathcal{J}$ is an open neighborhood of φ_1^*. Moreover, the operator

$$\mathbf{B}'(\chi, \epsilon; \varphi^*) \,:\, X_1 \to Y \times Y_*,$$

is invertible for every small enough $\epsilon > 0$ and for every $\varphi^* < \frac{c^2}{gH}\varphi_1^*$.

Proof. The first part of the lemma follows by a trivial generalization to the operator (4.5) of the arguments of the previous section. Let us now consider the operator $\mathbf{B}'(\chi, 0; \varphi^*)$; by Remark 3.6, the invertibility of this operator is equivalent to the extension of Theorem 3.1 and Corollary 3.3 to the problem obtained by adding to (3.6)–(3.8) the condition

$$\tilde{\eta} = \kappa \qquad \text{for} \quad \sigma = -cH/\varphi_1^*, \ \rho \in \mathbf{R},$$

with $\kappa \in Y_*$. One verifies that the proofs of Theorem 3.1 and Corollary 3.3 can be suitably generalized to the above problem, provided the condition $\nu^* cH/\varphi_1^* < 1$ holds; but this is equivalent to $\varphi^* < \frac{c^2}{gH}\varphi_1^*$. Now, since by (3.13) and (3.15), $\mathbf{B}'(\chi, \epsilon; \varphi^*) = \mathbf{B}'(\chi, 0; \varphi^*) + \mathcal{O}(\epsilon)$, the lemma follows. ∎

We can now prove the following.

PROPOSITION 4.4. *Let $\varphi^* > 0$ belong to the interval defined in the previous lemma. Then, for any $\rho \in \mathbf{R}$ and for every small $\epsilon \geq 0$, the map*

$$\varphi^* \mapsto \chi^\epsilon(\rho; \varphi^*) \tag{4.6}$$

is continuous.

Proof. Take $0 < \varphi_1^* \leq \varphi_2^*$ and, for the sake of brevity, define $\chi_1^\epsilon = \chi^\epsilon(\cdot; \varphi_1^*)$ $\chi_2^\epsilon = \chi^\epsilon(\cdot; \varphi_2^*)$. Note that the above solutions are defined in two different strips, A_1^*, A_2^*, with $A_2^* \subseteq A_1^*$; nevertheless, we can extend χ_2^ϵ to a strip A^* with $\varphi^* = \varphi_2^*/2$ by the Schwarz reflection principle (see, e.g., [5, Theorem 3.4]). Hence, by taking $\varphi_2^* \leq 2\varphi_1^*$, we can compare the two functions in the larger strip A_1^*. We may assume that χ_1^ϵ and χ_2^ϵ belong to $\mathcal{V}$; define now $\chi_\lambda^\epsilon = \lambda\chi_1^\epsilon + (1-\lambda)\chi_2^\epsilon$ and consider the identity

$$\int_0^1 \mathbf{B}'(\chi_\lambda, \epsilon; \varphi_2^*)(\chi_1^\epsilon - \chi_2^\epsilon) d\lambda = \mathbf{B}(\chi_1^\epsilon, \epsilon; \varphi_2^*) - \mathbf{B}(\chi_2^\epsilon, \epsilon; \varphi_2^*), \tag{4.7}$$

where $\mathbf{B}$ is the operator defined in $W_p^2(A_1^*)$ by (4.5). By Lemma 4.3, it follows that the operator at the left-hand side is invertible for (ϵ, φ_2^*) in a sufficiently small neighborhood of $(0, \varphi_1^*)$. Then, we have

$$\chi_2^\epsilon = \chi_1^\epsilon - \left[\int_0^1 \mathbf{B}'(\chi_\lambda, \epsilon; \varphi_2^*) d\lambda \right]^{-1} [\mathbf{B}(\chi_1^\epsilon, \epsilon; \varphi_2^*) - \mathbf{B}(\chi_2^\epsilon, \epsilon; \varphi_2^*)]. \tag{4.8}$$

Letting now $\varphi_2^* \to \varphi_1^*$, we get

$$\mathbf{B}(\chi_1^\epsilon, \epsilon; \varphi_2^*) \to \mathbf{B}(\chi_1^\epsilon, \epsilon; \varphi_1^*) = 0,$$

in $Y \times Y_*$. Furthermore, we note that

$$\mathbf{B}(\chi_2^\epsilon, \epsilon; \varphi_2^*) = \left(0, 0, \eta_2^\epsilon|_{\sigma=-cH/\varphi_1^*}\right).$$

By recalling that η_2^ϵ is holomorphic and vanishing for $\sigma = -cH/\varphi_2^*$, we can write

$$\eta_2^\epsilon(\cdot, -cH/\varphi_1^*) = cH\left(\frac{1}{\varphi_2^*} - \frac{1}{\varphi_1^*}\right)\partial_\sigma \eta_2^\epsilon(\cdot, -cH/\bar\varphi^*),$$

where $\varphi_1^* < \bar\varphi^* < \varphi_2^*$. Now, since η_2 has been extended by reflection, by the discussion in the appendix we obtain that the function $\partial_\sigma \eta_2(\cdot, -cH/\bar\varphi^*)$ is uniformly bounded in Y_* with respect to $\bar\varphi^*$. Then, we also obtain

$$\lim_{\varphi_2^* \to \varphi_1^*} \mathbf{B}(\chi_2^\epsilon, \epsilon; \varphi_2^*) = 0,$$

in $Y \times Y_*$. By (4.8), we conclude $\chi_2^\epsilon \to \chi_1^\epsilon$ in $\mathcal{V}$. In particular, we have the continuity of the map (4.6). □

Before discussing the continuity equation, we still have to specify an additional assumption on the function f. To do so, we observe that the norm in $W_p^2(A^*)$ of the solution $\tilde\eta(\cdot; \varphi^*)$ of the problem L, is bounded by a constant (depending on φ^*) times the $W_p^{2-\frac{1}{p}}$ norm of the datum $f'(\frac{\varphi^*}{c}\rho)$ on $(-1, 1)$; on the other hand, scaling back to the interval $(-\varphi^*/c, \varphi^*/c)$, the L_p norms of the kth derivative of the above function rescale by a factor $(\varphi^*/c)^{k-1/p}$. Then, for every $\varphi^* > 0$, there exists $C = C(\varphi^*)$ such that

$$||\tilde\chi(\cdot; \varphi^*)||_{W_p^2(A^*)} \le C||f'||_{W_p^{2-\frac{1}{p}}(-\varphi^*/c, \varphi^*/c)}. \tag{4.9}$$

In particular, the quantity $|\tilde\xi(1, 0; \varphi^*)|$ is bounded as above. Then, by recalling the definition of the point x_0 given in Assumption (c) of the introduction, we further assume the following.

Assumption F. *There exists φ_0^*, with $0 < \varphi_0^* < cx_0$, such that*

$$|\tilde\xi(1, 0; \varphi_0^*)| < \frac{g}{c^2}|f(\varphi_0^*/c)|. \tag{4.10}$$

As it follows from the bound (4.9), a sufficient condition for (4.10) is that $f'(x)$ is small enough in a neighborhood of the origin, where $f(x)$ is bounded away from zero.

Let us now consider the continuity equation (2.16); by using (2.20) and the notations of this section, we can write it in the form

$$\frac{\varphi^*}{c}\int_1^{+\infty} \eta^\epsilon(s, 0; \varphi^*)ds = f\left(\frac{\varphi^*}{c}\int_0^1 [1 + \epsilon\xi^\epsilon(s, 0; \varphi^*)]ds\right). \tag{4.11}$$

We can now state the following.

THEOREM 4.5. *Take f satisfying the assumptions of Theorem* 3.5 *and such that Assumption* F *holds. Then,* (4.11) *has a solution in the interval* $(0, cx_0)$.

Proof. By Proposition 4.4, the function

$$\mathcal{F}(\varphi^*) = \frac{\varphi^*}{c}\int_1^{+\infty} \eta^\epsilon(s,0;\varphi^*)ds - f\left(\frac{\varphi^*}{c}\int_0^1 [1+\epsilon\xi^\epsilon(s,0;\varphi^*)]ds\right)$$

is continuous in $(0, cx_0]$ for every ϵ in a neighborhood of zero. Moreover, by Theorem 4.1 and by noting that the right-hand side of (4.11) converges to $f(\varphi^*/c)$ for $\epsilon \to 0$, we find that $\mathcal{F}(cx_0) < 0$ for small enough ϵ. Now, by using the Bernoulli condition $B^F = 0$, we can also write

$$\mathcal{F}(\varphi^*) = \frac{c^2}{2g}\frac{2\xi^\epsilon(1,0;\varphi^*) + \epsilon[\xi^\epsilon(1,0;\varphi^*)^2 + \eta^\epsilon(1,0;\varphi^*)^2]}{|1+\epsilon\chi^\epsilon(1,0;\varphi^*)|^2}$$

$$-f\left(\frac{\varphi^*}{c}\int_0^1 [1+\epsilon\xi^\epsilon(s,0;\varphi^*)]ds\right). \tag{4.12}$$

By Assumption F and recalling that $f(x)$ is negative for $x < x_0$, we have that the right-hand side of (4.12) is strictly positive at φ_0^* for $\epsilon = 0$. Now, since $\chi^\epsilon(1,0;\varphi^*) = \tilde{\chi}(1,0;\varphi^*) + \mathcal{O}(\epsilon)$ we get $\mathcal{F}(\varphi_0^*) > 0$ for every small enough ϵ. Then, the theorem follows. ∎

REMARK 4.6. *By the above theorem, by Theorem* 3.8*, and by relations* (2.17)*,* (2.19)*,* (2.20)*, it follows that there exists a function* $\Omega = U - iV$ *holomorphic in the strip* A_H *defined by* (2.4) *and a positive number* $\varphi^* < cx_0$ *such that* (2.12)–(2.16) *hold for every small enough* $\epsilon > 0$*. Moreover, the map* (2.8) *is one-to-one between* A_H *and the domain* S^* *in the physical plane defined by* (1.2)*, with the free boundary* $h(x)$ *given in parametric form by* $x = x(\varphi, 0)$*,* $y = y(\varphi, 0)$*,* $|\varphi| > \varphi^*$*. Finally, the point* $x^* = x(\varphi^*, 0)$*, the function* $h(x)$*, and the function* $\omega(z)$ *given by* (2.1) *satisfy the conditions* (1.3)–(1.9)*. Notice that these results hold under rather mild geometrical requirements on the curve* $y = f(x)$*; nevertheless, we still have to satisfy the last condition* (1.10) *in the physical plane. To do so, we assume in addition that the function* f *is* convex.

PROPOSITION 4.7. *Let* f *be a convex function satisfying the assumptions of Theorem* 3.5 *and Assumption* F *above; let* $h(x)$ *be defined as in Remark* 4.6*. Then, for every small enough positive* ϵ*, we have* $h(x) < f(x)$ *for* $x \in I\backslash[-x^*, x^*]$*.*

Proof. By symmetry and by the convexity of f, it is enough to prove that $h'(x) \le f'(x^*)$ for $x \ge x^*$. This is equivalent to the condition

$$-\frac{\eta^\epsilon(\rho,0)}{1+\epsilon\xi^\epsilon(\rho,0)} \le f'(x^*) \qquad \text{for} \quad \rho \ge 1,$$

where, as before, $\chi^\epsilon = \xi^\epsilon - i\eta^\epsilon$ is the solution in the (rescaled) hodograph plane. By defining

$$\psi^\epsilon(\rho,\sigma) = \eta^\epsilon(\rho,\sigma) + \epsilon f'(x^*)\xi^\epsilon(\rho,\sigma), \tag{4.13}$$

we get the equivalent condition

$$\psi^\epsilon(\rho,0) \ge -f'(x^*) \qquad \text{for} \quad \rho \ge 1. \tag{4.14}$$

Assume that (4.14) does not hold and set

$$\inf_{\rho\ge 1} \psi^\epsilon(\rho,0) \equiv -m^* < -f'(x^*).$$

Since ψ^ϵ is continuous, vanishing at infinity, and $\psi^\epsilon(1,0) = -f'(x^*)$, it must be $m^* = \psi^\epsilon(\rho^*, 0)$, with $\rho^* = \rho^*(\epsilon) > 1$ uniformly with respect to ϵ. Let us define the comparison function

$$W^* = -(m^* + \epsilon K^*)\frac{\varphi^*}{cH}\,\sigma - m^*,$$

where $K^* < f'(x^*)\inf_{A^*}\xi^\epsilon$. Then, recalling the conditions satisfied by η^ϵ, we get $\psi^\epsilon > W^*$ for $\sigma = -cH/\varphi^*$ and on the segment $\{0\}\times[-cH/\varphi^*, 0]$ for every suitably small ϵ. Furthermore, for $\sigma = 0$ and $0 < \rho < 1$ we get, by using the condition $B^I(\chi^\epsilon, \epsilon) = 0$ and the convexity of f,

$$\psi^\epsilon(\rho,0) = -f'(x(\rho,0))[1+\epsilon\xi^\epsilon(\rho,0)] + \epsilon f'(x^*)\xi^\epsilon(\rho,0)$$

$$= -f'(x^*) + [f'(x^*) - f'(x(\rho,0))][1+\epsilon\xi^\epsilon(\rho,0)] \geq -f'(x^*).$$

Hence, we also have $\psi^\epsilon(\rho,0) \geq W^*(\rho,0) = -m^*$, for $\rho \geq 0$, with the equal sign holding at ρ^*. It follows that $\psi^\epsilon \geq W^*$ in the half strip $[0,+\infty]\times[-cH/\varphi^*, 0]$ and that

$$\frac{\partial\psi^\epsilon}{\partial\sigma}(\rho^*,0) < -(m^* + \epsilon K^*)\frac{\varphi^*}{cH}. \tag{4.15}$$

Now, since ρ^* is a minimum for $\psi^\epsilon(\rho,0)$, we also get $\eta^\epsilon_\rho + \epsilon f'(x^*)\xi^\epsilon_\rho = 0$ at $(\rho^*,0)$, and therefore

$$\xi^\epsilon_\sigma(\rho^*,0) = \epsilon f'(x^*)\eta^\epsilon_\sigma(\rho^*,0).$$

By inserting into (4.15), we obtain

$$\eta^\epsilon_\sigma(\rho^*,0) < -\frac{m^* + \epsilon K^*}{1+\epsilon^2 f'(x^*)^2}\,\frac{\varphi^*}{cH}. \tag{4.16}$$

Notice that $\xi^\epsilon_\rho = -\eta^\epsilon_\sigma > 0$ and $\eta^\epsilon_\rho < 0$ at $(\rho^*,0)$. Moreover, by the Bernoulli condition, we have at the same point

$$\eta^\epsilon_\sigma = \frac{g\varphi^*}{c^3}\,\frac{|1+\epsilon\chi^\epsilon|^4 + \epsilon\eta^\epsilon_\rho}{1+\epsilon\xi^\epsilon}\,\eta^\epsilon$$

$$= -\frac{g\varphi^*}{c^3}\,\frac{|1+\epsilon\chi^\epsilon|^4 + \epsilon\eta^\epsilon_\rho}{1+\epsilon\xi^\epsilon}\,[m^* + \epsilon f'(x^*)\xi^\epsilon]. \tag{4.17}$$

From (4.16), (4.17), and the above remark on η^ϵ_ρ, we get the condition

$$\frac{|1+\epsilon\chi^\epsilon|^4}{1+\epsilon\xi^\epsilon}\,[m^* + \epsilon f'(x^*)\xi^\epsilon] > \frac{c^2}{gH}\,\frac{m^* + \epsilon K^*}{1+\epsilon^2 f'(x^*)^2} \tag{4.18}$$

at the point $(\rho^*, 0)$. Clearly, (4.18) contradicts the condition of supercritical velocity for ϵ small enough, so that the proposition is proven. □

Summing up the discussion of the present section, we can finally state the following.

THEOREM 4.8. *Let f be a convex function satisfying the assumptions of Theorem* 3.5 *and Assumption* F. *Then, for every small enough $\epsilon > 0$, there exists a real number*

$x^* \in (0, x_0)$, a real symmetric function $h(x)$ on $\mathbf{R}\backslash[-x^*, x^*]$, and a complex function ω holomorphic in the domain S^* defined by (1.2) *and satisfying $\omega(-x+iy) = \bar{\omega}(x+iy)$, such that the conditions* (1.3)–(1.10) *hold. Moreover, the function h is negative and monotone increasing for $x > x^*$.*

REMARK 4.9. *We remark that the qualitative shape of the free boundary given by the above theorem agrees with the numerical results obtained in* [8].

REMARK 4.10. *We point out that in the solution above the free boundary and the cylinder profile form a single smooth ($\mathcal{C}^1$) streamline. Nevertheless, it seems reasonable to conjecture that there exist other solutions which are less regular at the points $(\pm x^*, f(\pm x^*))$, where the free boundary meets the hull. This conjecture is supported by the existence of nonvariational solutions of the linearized problem, as shown in* [5, section 4].

Appendix. Proof of Corollary 3.3. Let us recall that we are seeking a function $\tilde{\chi} = \tilde{\xi} - i\tilde{\eta} \in W_p^2(A^*)$, holomorphic in A^*, satisfying the boundary conditions (3.7), (3.9), (3.10), and the properties

$$\tilde{\chi}(-\rho, \sigma) = \overline{\tilde{\chi}(\rho, \sigma)}, \tag{A.1}$$

$$\tilde{\chi}|_{\sigma=0} \in \mathcal{C}^{1,\alpha}(\mathbf{R}\backslash(-\rho_0, \rho_0)),$$

$$\sup_{A^*} e^{\lambda^*|\rho|}|\tilde{\chi}(\rho, \sigma)| < \infty, \tag{A.2}$$

$$\sup_{|\rho|\geq\rho_0} e^{\lambda^*|\rho|}|\partial_\rho \tilde{\xi}(\rho, 0)| < \infty. \tag{A.3}$$

Proof. We first note that for data l and k antisymmetric, the solution $\tilde{\eta}$ given by Theorem 3.5 necessarily satisfies $\eta(-\rho, \sigma) = -\eta(\rho, \sigma)$; therefore, the harmonic conjugate (defined up to an arbitrary constant) is symmetric with respect to ρ, so that (A.1) holds.

By standard continuation results, we may assume that the datum l belongs to $W_p^{1-\frac{1}{p}}(\mathbf{R})$. Then, the solution of the problem (3.6)–(3.9) can be written in the form $\tilde{\eta} = \eta_0 + \eta_1$, where η_0, η_1 are harmonic in A^*, vanish at $\sigma = -\frac{cH}{\varphi^*}$, and satisfy the boundary conditions

$$\partial_\sigma \eta_0 - \nu^* \eta_0 = 0 \qquad \text{for} \quad \sigma = 0, \ \ |\rho| > 1, \tag{A.4}$$

$$\eta_0 = k - \eta_1 \qquad \text{for} \quad \sigma = 0, \ \ |\rho| < 1, \tag{A.5}$$

$$\partial_\sigma \eta_1 - \nu^* \eta_1 = l \qquad \text{for} \quad \sigma = 0, \ \ \rho \in \mathbf{R}. \tag{A.6}$$

We observe that, if η_1 is known, the problem for η_0 is similar to problem L; hence, by the results of [5], the bounds (A.2)–(A.3) hold for the holomorphic function $\chi_0 = \xi_0 - i\eta_0$ (where ξ_0 is the harmonic conjugate of η_0 vanishing at infinity). Thus, we are reduced to proving the bounds for the function η_1 satisfying (A.6) (and for the harmonic conjugate ξ_1). Let us define $H^* = \frac{cH}{\varphi^*}$; by elementary calculations, η_1 has the representation

$$\eta_1(\rho, \sigma) = \frac{1}{2\pi} \int_{\mathbf{R}} e^{ip\rho} \hat{K}_\sigma(p) \hat{l}(p) dp, \tag{A.7}$$

where

$$\hat{K}_\sigma(p) = \frac{\sinh[p(\sigma + H^*)]}{p\cosh(pH^*) - \nu^* \sinh(pH^*)}$$

and $\hat{l}(p)$ is the Fourier transform of l. We point out that the function $\hat{K}_\sigma$ is not singular since the equation $\nu^* \tanh(pH^*) = p$ has only the real solution $p = 0$ for $\nu^* H^* < 1$. We further note that the integral (A.7) is convergent also for $\sigma = 0$. In fact, by Sobolev embedding, $l \in L^q(\mathbf{R})$ for $1 < q < 2$ and therefore, by the Hausdorff–Young theorem, $\hat{l} \in L^p(\mathbf{R})$ for every $p > 2$; thus, the product $\hat{K}_\sigma(p)\hat{l}(p)$ is integrable by Hölder inequality. In particular, we recover the continuity of η_1 up to the boundary $\sigma = 0$ (see Remark 3.2). By the convolution theorem we have

$$\eta_1(\rho, \sigma) = \int_{\mathbf{R}} K_\sigma(\rho - \rho') l(\rho') d\rho', \tag{A.8}$$

where

$$K_\sigma(\rho) = \frac{1}{2\pi} \int_{\mathbf{R}} e^{ip\rho} \frac{\sinh[p(\sigma + H^*)]}{p\cosh(pH^*) - \nu^* \sinh(pH^*)} dp. \tag{A.9}$$

Note that the function $\rho \mapsto K_\sigma(\rho)$ is smooth and rapidly decreasing for $\sigma < 0$ and belongs to $L^2(\mathbf{R})$ for $\sigma = 0$ (actually, to $L^p(\mathbf{R})$ for $1 \le p \le 2$). Moreover, $K_\sigma(-\rho) = K_\sigma(\rho)$.

For $|\rho| > 0$, we can evaluate (A.9) by complex plane integration and find

$$K_\sigma(\rho) = \sum_{n=1}^{\infty} c_n(\sigma) e^{-\lambda_n^* |\rho|}, \tag{A.10}$$

where

$$c_n(\sigma) = \frac{\sin[\lambda_n^*(\sigma + H^*)]}{(1 - \nu^* H^*)\cos(\lambda_n^* H^*) - \lambda_n^* H^* \sin(\lambda_n^* H^*)}$$

and λ_n^* are the positive solutions of the equation

$$\tan(\lambda H^*) = \frac{\lambda}{\nu^*}.$$

Note that $\lambda_n^* H^* \approx (n - 1/2)\pi$ for large n, so that $c_n(0) \sim -1/n\pi$. From (A.10) we get the estimate

$$|K_\sigma(\rho)| \le C e^{-\lambda_1^* |\rho|} \tag{A.11}$$

for $|\rho| \ge \delta > 0$, with C independent of σ.

We can now prove the bounds (A.2)–(A.3) for the holomorphic function $\chi_1 = \xi_1 - i\eta_1$. We set $I_0 = (-\rho_0, \rho_0)$, $I_{\rho,\delta} = (\rho - \delta, \rho + \delta)$; then, by the representation (A.8), the estimate (A.11), and the decaying property of l, we obtain for $|\rho| > \rho_0 + \delta$

$$|\eta_1(\rho, \sigma)| \le \int_{I_{\rho,\delta}} |K_\sigma(\rho - \rho')|\, |l(\rho')|\, d\rho' + \int_{I_0} |K_\sigma(\rho - \rho')|\, |l(\rho')|\, d\rho'$$

$$+ \int_{\mathbf{R}/\{I_{\rho,\delta} \cup I_0\}} |K_\sigma(\rho - \rho')|\, |l(\rho')|\, d\rho'$$

$$\leq C\left\{(||K_\sigma||_{L^2}+||l||_{L^p})e^{-\lambda_1^*|\rho|}+\int_{\mathbf{R}/\{I_{\rho,\delta}\cup I_0\}} e^{-\lambda_1^*(|\rho-\rho'|-|\rho'|)}d\rho'\right\}$$

$$\leq Ce^{-\lambda_1^*|\rho|}, \tag{A.12}$$

with C independent of σ.

By (A.8)–(A.9), we see that the same bound holds for every derivative of η_1 if $\sigma \leq \sigma_0 < 0$. On the other hand, by the condition (A.6) and the assumption on l, the estimate for $\partial_\sigma \eta_1$ extends to $\sigma = 0$ for $|\rho| \geq \rho_0$; the same holds for the function $\partial_\rho \xi_1$ by the Cauchy–Riemann relations, so that (A.3) holds. Furthermore, we have that $\partial_\rho \xi_1(\rho, 0)$ is in $L^1(\mathbf{R})$ (for $|\rho| < \rho_0$ it is in L^p with $p > 1$) and we can write

$$\xi_1(\rho,0)=\int_{-\infty}^{\rho}\partial_\rho\xi_1(\rho,0)=-\int_{\rho}^{-\infty}\partial_\rho\xi_1(\rho,0),$$

where the last equality follows by the symmetry of ξ_1. From the above discussion, it follows also that ξ_1 satisfies the bound (A.12). Then, (A.2) is proved.

It remains to prove the regularity of the traces $\eta_1(\rho, 0)$ and $\xi_1(\rho, 0)$ for $|\rho| \geq \rho_0$. By our assumption on h, by Remark 3.2, and recalling (A.6), we get that the harmonic function η_1 has a Neumann datum in $\mathcal{C}^{0,\alpha}(\mathbf{R}\backslash(-\rho_0, \rho_0))$ on the boundary $\sigma = 0$. Then, by the previous bounds on ξ_1, η_1 and by standard Hölder estimates, $\chi_1|_{\sigma=0} \in \mathcal{C}^{1,\alpha}(\mathbf{R}\backslash(-\rho_0, \rho_0))$ □

REMARK A.1. *The boundedness of the trace function $\partial_\rho \eta_1|_{\sigma=0}$ for $|\rho| \geq \rho_0$ can also be proved from the representation* (A.8). *Actually, for $\sigma < 0$ we get*

$$\partial_\rho\eta_1(\rho,\sigma)=\int_{\mathbf{R}}K'_\sigma(\rho-\rho')l(\rho')d\rho',$$

where $K'_\sigma(\rho)$ is the inverse Fourier transform of the function $ip\hat{K}_\sigma(p)$. We point out that $K'_\sigma(\rho)$ is an odd function in the Schwartz space for $\sigma < 0$; moreover, by the same arguments as before, we get the estimate

$$|K'_\sigma(\rho)| \leq C\frac{e^{-\lambda_1^*|\rho|}}{1-e^{-|\rho|/H^*}}$$

for every $\rho \neq 0$, with C independent of σ. Thus, for $|\rho| > \rho_0 + \delta$ we have

$$|\partial_\rho\eta_1(\rho,\sigma)| \leq \int_{I_{\rho,\delta}}|K'_\sigma(\rho-\rho')|\,|l(\rho)-l(\rho')|\,d\rho'+\int_{\mathbf{R}/\{I_{\rho,\delta}\}}|K_\sigma(\rho-\rho')|\,|l(\rho')|\,d\rho'$$

$$\leq C\left\{\int_{-\delta}^{\delta}\frac{e^{-\lambda_1^*|r|}}{1-e^{-|r|/H^*}}|r|^\alpha dr+\frac{e^{-\lambda_1^*|\delta|}}{1-e^{-|\delta|/H^*}}||l||_{L^1(\mathbf{R})}\right\},$$

where C is independent of σ. The result now follows by the above estimate and by the continuity of the trace $\partial_\rho \eta_1(\rho, 0)$ on $\mathbf{R}\backslash(-\rho_0, \rho_0)$.

REMARK A.2. *It is worthwhile to remark that the above proof can be easily generalized to the case of a datum $l \in \mathcal{C}^{k,\alpha}(\mathbf{R}\backslash(-\rho_0, \rho_0))$, with arbitrary k. As a consequence, the solution of the nonlinear problem is $\mathcal{C}^{k+1,\alpha}$ up to the boundary for $|\rho| > \rho_0$. Hence, by the arbitrariness of k and ρ_0, we conclude that the solution χ^ϵ is actually smooth up to the boundary $\sigma = 0$ for $|\rho| > 1$.*

Acknowledgment. Useful discussions and comments by Prof. Avner Friedman are gratefully acknowledged.

REFERENCES

[1] N. G. Kuznetsov and V. G. Maz'ya, *Unique solvability of the plane Neumann-Kelvin problem*, Math. USSR-Sb., 63 (1989), pp. 425–446.

[2] N. G. Kuznetsov, *On uniqueness and solvability in the linearized two-dimensional problem of a supercritical stream about a surface-piercing body*, Proc. Roy. Soc. London Ser. A, 450 (1995), pp. 233–253.

[3] F. Ursell, *Mathematical notes on the two-dimensional Kelvin-Neumann problem*, in Proceedings of the 13th Symposium on Naval Hydrodynamics, Tokyo 1980, Shipbuilding Res. Assoc. Japan, Tokyo, Japan, 1981, pp. 145–151.

[4] C. D. Pagani and D. Pierotti, *Exact solution of the wave resistance problem for a submerged cylinder.* II. *The non-linear problem*, Arch. Rational Mech. Anal., 149 (1999), pp. 289–327.

[5] C. D. Pagani and D. Pierotti, *The Neumann-Kelvin problem for a beam*, J. Math. Anal. Appl., 240 (1999), pp. 60–79.

[6] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Pitman, London, 1985.

[7] W. Craig and P. Sternberg, *Comparison principles for free-surface flows with gravity*, J. Fluid Mech., 230 (1991), pp. 231–243.

[8] J. Asavanant and J. M. Vanden-Broeck, *Free surface flows past a surface-piercing object of finite length*, J. Fluid Mech., 273 (1994), pp. 109–124.

SIAM J. MATH. ANAL.
Vol. 32, No. 2, pp. 235–260

INTERFACE DEVELOPMENT AND LOCAL SOLUTIONS TO REACTION-DIFFUSION EQUATIONS*

UGUR G. ABDULLA† AND JOHN R. KING‡

Abstract. The evolution of interfaces and the local behavior of solutions near the interface in problems for one-dimensional reaction-diffusion equations are studied. In all cases explicit formulae for the interface, with accuracy up to constant coefficients, are found together with the local solution. The methods used are matched asymptotic expansions for preliminary formal results, and rescaling and a barrier technique for rigorous proof, using special comparison theorems in irregular domains.

Key words. reaction-diffusion equations, evolution of interfaces, nonlinear degenerate parabolic equations, local solutions

AMS subject classifications. 35K55, 35K65

PII. S003614109732986X

1. Introduction. We consider the Cauchy problem (CP) for the reaction-diffusion equation

$$\mathrm{L}u \equiv u_t - (u^m)_{xx} + bu^\beta = 0, \quad x \in \mathbb{R}, \quad 0 < t < T, \tag{1.1}$$

with

$$u(x,0) = u_0(x), \qquad x \in \mathbb{R}, \tag{1.2}$$

where $m > 1,\ b \in \mathbb{R},\ \beta > 0,\ 0 < T \le +\infty$, and u_0 is nonnegative and continuous. Furthermore, we will suppose that $b > 0$ if $\beta < 1$, and b is arbitrary if $\beta \ge 1$ (see Remark 1.1). Such reaction-diffusion equations are widely used models for various physical, chemical, and biological problems involving diffusion with a source or with absorption, such as occurs, for instance, in filtration in porous media, flow of a chemically reacting fluid from a flat surface, evolution of biological populations, etc.

The solution to (1.1), (1.2) may have one or several interfaces separating the regions where $u = 0$ and where $u > 0$. In this paper we are interested in the small-time evolution of interfaces and in the local structure of solutions near the interface. Since (1.1) is invariant under the transformations $x \to -x,\ x \to x + c,\ c \in \mathbb{R}$, without loss of generality we will investigate the case when $\eta(0) = 0$, where

$$\eta(t) = \sup\{x :\ u(x,t) > 0\}.$$

More precisely, we are interested in the short-time behavior of the interface function $\eta(t)$ and local solution near the interface. Furthermore, unless otherwise stated, we shall assume that

$$u_0 \sim C(-x)_+^\alpha \text{ as } x \to 0- \quad \text{for some } C > 0,\ \alpha > 0. \tag{1.3}$$

*Received by the editors November 11, 1997; accepted for publication (in revised form) June 23, 1999; published electronically June 27, 2000.
http://www.siam.org/journals/sima/32-2/32986.html
†Faculty of Applied Mathematics and Cybernetics, Baku State University, Baku 370148, Azerbaijan. Current address: Max-Planck Institute for Mathematics in the Sciences, Inselstrasse 22, Leipzig 04103, Germany (Ugur.Abdulla@mis.mpg.de).
‡School of Mathematical Sciences: Department of Theoretical Mechanics, University of Nottingham, Nottingham NG7 2RD UK (john.king@nottingham.ac.uk).

We present a full description of the small-time behavior of $\eta(t)$ and local solution near $\eta(t)$ for all relevant values of parameters m, b, β, C, and α (see Remark 1.1). The behavior of u_0 as $x \to -\infty$ has no influence on our results. Accordingly, we may suppose that u_0 either is bounded or satisfies some restriction on its growth rate as $x \to -\infty$ which is suitable for existence, uniqueness, and comparison results (see section 3). In addition, in some cases we shall consider the special case

$$u_0(x) = C(-x)_+^{\alpha}, \quad x \in \mathbb{R}, \tag{1.4}$$

namely, when the solution to the problem (1.1), (1.4) is of self-similar form. In these cases our estimations on $\eta(t)$ and the local solution near $\eta(t)$ will be global in time.

Initial development of interfaces in problems for (1.1) have been studied by many authors [2, 4, 6, 7, 8, 10, 13, 17, 18, 19, 22, 25, 27, 28, 29, 30, 31, 33, 34, 35, 36]. All values of m, b, β, C, and α for which the interface initially either shrinks, remains stationary, or expands are known due to these papers. As to local estimations of both interface and solution there is a complete picture only in the case of the semilinear equation ($m = 1$), given in [17, 18]. Accordingly, we are not directly interested in the case of $m = 1$, although when our results contain the semilinear equation as a particular case, it will be mentioned. It should be noted that sometimes this may not be the case. For instance, if $b > 0,\ 0 < \beta < 1,\ \alpha < 2/(m - \beta)$, then the interface initially expands and

$$\eta(t) \sim C_1 t^{1/(2-\alpha(m-1))} \quad \text{as } t \to 0+.$$

Formally, as $m \to 1$ this estimate yields

$$\eta(t) = O(t^{\frac{1}{2}}) \quad \text{as } t \to 0+,$$

while as a matter of fact from [18] it follows that if $m = 1$, then

$$\eta(t) \sim C_2(t \log 1/t)^{\frac{1}{2}} \quad \text{as } \ t \to 0+$$

($C_1,\ C_2$ are positive constants), so that the case $m = 1$ is in some respects a singular limit.

Many of the results of this paper have first been formally established using matched asymptotic expansions (JRK). For rigorous proof rescaling and barrier techniques that use special comparison theorems in irregular domains have been used (UGA). The latter is the main difference of the methods of our paper from those of previous papers (including [17, 18]). Similar barrier techniques using standard comparison theorems in cylindrical domains have been applied earlier [2] to the same problem. As a result, in [2] explicit formulae for the interface and for the local solution have been derived, but only in some particular cases when the small-time behavior of the solution has a uniform character near the interface (see also [4]). In a many cases, however, the behavior is nonuniform in the sense of singular perturbation theory, the dominant balance as $t \to 0+$ between the terms in (1.1) on curves which approach the boundary of the support on the initial line depending on how they do so. In order to apply a barrier technique to these cases as well, it is necessary to investigate mathematical problems of the general theory (existence, uniqueness, and comparison results) of initial-boundary value problems for reaction-diffusion equations in noncylindrical domains with boundary curves which may be nonsmooth and characteristic at the initial moment. These issues have been investigated in a recent

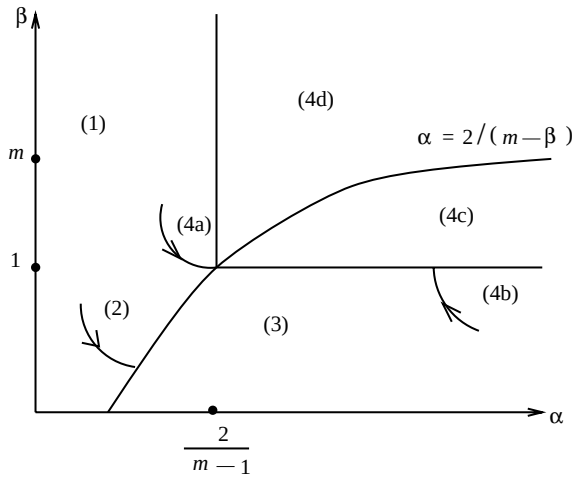


FIG. 1. *Classification of different cases in the* (α, β) *plane for interface development in problem* (1.1)–(1.4).

paper [3]. The comparison theorems from [3] are widely used in the proof of the main results of this paper.

The organization of the paper is as follows: In section 2 we outline the main results. In section 3 we then apply scale of variables methods for some preliminary estimations which are necessary for using our barrier technique. Finally in section 4 we prove the results of section 2.

To avoid difficulties for the reader we give explicit values of some of the constants which appear in sections 2 and 4 (the case I(2)) in the appendix.

REMARK 1.1. *We shall not consider in this paper the case* $b < 0,\ 0 < \beta < 1$ *when, for some range of* m *and* β*, nonuniqueness of the solution to* (1.1), (1.2) *is possible. The Cauchy problem* (1.1), (1.2) *in this range of parameters has been extensively investigated in* [33, 34]. *Existence and boundary regularity results for Cauchy–Dirichlet and Dirichlet problems for equation* (1.1) *in noncylindrical domains with nonsmooth boundaries which are applicable to this case have been established in* [3]*, but in order to prove similar results for the problem of the evolution of interfaces in this case it would be important first to establish special comparison results in such domains as well. It should also be pointed out that by using the results of* [3]*, a similar classification of the evolution of (possible) interfaces and local solutions may be given in the fast diffusion case* $(0 < m < 1)$. *We shall address these issues in a subsequent paper.*

2. Description of main results. We divide the results into the two different subcases:

(I) $b \neq 0$ (either $b > 0,\ \beta > 0$ or $b < 0,\ \beta \geq 1$) and $m > 1$;

(II) $b = 0$.

(I) In this case there are four different subcases, as shown in Figure 1. (In view of our assumptions, the case $b < 0$ relates to the part of the (α, β) plane with $\beta \geq 1$.)

(1) Suppose that $\alpha < 2/(m - \min\{1, \beta\})$. In this case the interface initially expands and

$$\eta(t) \sim \xi_* t^{1/(2-\alpha(m-1))} \quad \text{as} \quad t \to 0+, \tag{2.1}$$

where

$$\xi_* = C^{\frac{m-1}{2-\alpha(m-1)}}\xi'_*, \tag{2.2}$$

and ξ'_* is a positive number depending on m and α only (see Lemma 3.1). For arbitrary $\rho < \xi_*$, there exists a positive number $f(\rho)$ depending on C, m, and α such that

$$u(x,t) \sim f(\rho)t^{\alpha/(2-\alpha(m-1))} \quad \text{as} \quad t \to 0+ \tag{2.3}$$

along the curve $x = \xi_\rho(t) = \rho t^{1/(2-\alpha(m-1))}$. Actually, f is a self-similar solution to the problem (1.1), (1.4) with $b = 0$ (see Lemma 3.1) and

$$f(\rho) \ = C^{2/(2-\alpha(m-1))} f_0\Big(C^{(m-1)/(\alpha(m-1)-2)}\rho\Big), \tag{2.4a}$$

$$f_0(\rho) = \omega(\rho, 1), \qquad \xi'_* = \sup\Big\{x : \ f_0(\rho) > 0\Big\} > 0, \tag{2.4b}$$

where ω is a solution of the CP (1.1), (1.4) with $b = 0, C = 1$. Lower and upper estimations for f are given in (2.27). We also have that

$$\xi'_* = A_0^{\frac{m-1}{2}} \left[\frac{m(2-\alpha(m-1))}{m-1}\right]^{\frac{1}{2}} \xi''_*, \tag{2.5}$$

where $A_0 = \omega(0,1)$ and ξ''_* is some number belonging to the segment $[\xi_1, \xi_2]$, where

$$\begin{aligned} &\xi_1 = \Big(\alpha(m-1)\Big)^{-\frac{1}{2}}, \qquad \xi_2 = 1 && \text{if} \ \ (m-1)^{-1} \le \alpha < 2(m-1)^{-1}, \\ &\xi_1 = 1, && \xi_2 = \Big(\alpha(m-1)\Big)^{-\frac{1}{2}} \qquad \text{if} \ \ 0 < \alpha \le (m-1)^{-1}. \end{aligned} \tag{2.6}$$

In particular, if $\alpha = (m-1)^{-1}$ and $m > 2 - \min\{1, \beta\}$, then the explicit solution of the problem (1.1), (1.4) with $b = 0$ is given by (2.24) and we have

$$\xi_1 = \xi_2 = 1, \quad \xi'_* = m(m-1)^{-1}, \quad f_0(x) = \Big(\xi'_* - x\Big)_+^{1/(m-1)}, \tag{2.7}$$

where $(r)_+ = \max(r, 0)$.
The explicit formulae (2.1) and (2.3) mean that the local behavior of the interface and solution along $x = \xi_\rho(t)$ coincide with those of the problem (1.1), (1.4) with $b = 0$.

(2) Suppose that $b > 0, \ 0 < \beta < 1, \ \alpha = 2/(m-\beta)$ (here we describe the results for the case $m = 1$ as well). In this case the behavior of the interface depends on the constant C. The critical value is

$$C_* = \Big[|b|(m-\beta)^2 \Big/ \Big(2m(m+\beta)\Big)\Big]^{1/(m-\beta)}.$$

First, assume that u_0 is defined by (1.4). If $m + \beta = 2$, then the explicit solution to (1.1), (1.4) is

$$u(x,t) = C(\zeta_* t - x)_+^{1/(1-\beta)}, \qquad \zeta_* = b(1-\beta)C^{\beta-1}\left((C/C_*)^{m-\beta} - 1\right). \tag{2.8}$$

It has an expanding interface if $C > C_*$ and a shrinking interface if $0 < C < C_*$ and is a stationary solution if $C = C_*$.

Let $m + \beta \neq 2$. If $C = C_*$ then u_0 is a stationary solution to (1.1), (1.4). If $C \neq C_*$, then the solution to (1.1), (1.4) is of the self-similar form

$$u(x,t) = t^{1/(1-\beta)} f_1(\zeta), \qquad \zeta = x t^{-\frac{m-\beta}{2(1-\beta)}}, \tag{2.9}$$

$$\eta(t) \quad = \zeta_* t^{\frac{m-\beta}{2(1-\beta)}}, \qquad 0 \le t < +\infty. \tag{2.10}$$

If $C > C_*$ then the interface expands, $f_1(0) = A_1 > 0$ (see Lemma 3.3), and

$$C_1 t^{\frac{1}{1-\beta}} \Big(\zeta_1 - \zeta\Big)_+^{\mu} \le u \le C_2 t^{\frac{1}{1-\beta}} \Big(\zeta_2 - \zeta\Big)_+^{\frac{2}{m-\beta}}, \quad 0 \le x < +\infty, \quad 0 < t < +\infty, \tag{2.11}$$

where

$$\mu = (m-1)^{-1} \quad \text{if} \quad m + \beta > 2; \quad \mu = 2(m-\beta)^{-1} \quad \text{if} \quad 1 \le m < 2 - \beta,$$

which implies

$$\zeta_1 \le \zeta_* \le \zeta_2. \tag{2.12}$$

The right-hand side of (2.11) (respectively, (2.12)) may be replaced by $\bar{C}_2 t^{\frac{1}{1-\beta}} (\bar{\zeta}_2 - \zeta)_+^{\frac{1}{m-1}}$ (respectively, $\bar{\zeta}_2$); see the appendix for the description of all the relevant constants.

Let $m + \beta \neq 2$ and $0 < C < C_*$. Then the interface shrinks and if $m + \beta > 2$, then

$$\begin{aligned} &\left[C^{1-\beta}\Big(-x\Big)_+^{\frac{2(1-\beta)}{m-\beta}} - b(1-\beta)t\right]_+^{\frac{1}{1-\beta}} \le u \\ &\le \left[C^{1-\beta}\Big(-x\Big)_+^{\frac{2(1-\beta)}{m-\beta}} - b(1-\beta)\Big(1-(C/C_*)^{m-\beta}\Big)t\right]_+^{\frac{1}{1-\beta}}, \\ &\qquad\qquad x \in \mathbb{R}, \quad 0 \le t < +\infty, \end{aligned} \tag{2.13}$$

which also implies (2.12), where we replace ζ_1 (respectively, ζ_2) with

$$-C^{-\frac{m-\beta}{2}} \left(b(1-\beta)\right)^{\frac{m-\beta}{2(1-\beta)}}$$

$$\left(\text{respectively, } -C^{-\frac{m-\beta}{2}} \left(b(1-\beta)\left(1-(C/C_*)^{m-\beta}\right)\right)^{\frac{m-\beta}{2(1-\beta)}}\right).$$

However, if $1 \le m < 2 - \beta$, then

$$C_* \left(-\zeta_3 t^{\frac{m-\beta}{2(1-\beta)}} - x\right)_+^{\frac{2}{m-\beta}} \le u \le C_3 \left(-\zeta_4 t^{\frac{m-\beta}{2(1-\beta)}} - x\right)_+^{\frac{2}{m-\beta}}, \quad 0 \le t < +\infty, \tag{2.14}$$

where the left-hand side is valid for $x \ge -\ell_0 t^{\frac{m-\beta}{2(1-\beta)}}$, while the right-hand side is valid for $x \ge -\ell_1 t^{\frac{m-\beta}{2(1-\beta)}}$. From (2.14), (2.12) follows if we replace ζ_1 and ζ_2 with $-\zeta_3$ and $-\zeta_4$, respectively.

When $m+\beta \neq 2$, in general the precise value ζ_* can be found only by solving the similarity ODE $\mathcal{L}^0 f_1 = 0$ (see (4.4b) below) and calculating $\zeta_* = \sup\{\zeta : f_1(\zeta) > 0\}$. It may easily be shown that the described estimations (2.11)–(2.14), together with existence, uniqueness, and comparison results for the original Cauchy problem (1.1),(1.4) (see section 3), imply the unique solvability of the relevant boundary value problems for the similarity ODE, as well as the existence and uniqueness of ζ_*. Respective lower and upper bounds for ζ_* are given in (2.12).
Now assume that u_0 satisfies (1.3) with $\alpha = 2/(m-\beta)$. Then if $C \neq C_*$ we have

$$\eta(t) \sim \zeta_* t^{\frac{m-\beta}{2(1-\beta)}} \quad \text{as} \quad t \to 0+ \tag{2.15}$$

and for arbitrary $\rho < \zeta_*$

$$u(x,t) \sim f_1(\rho) t^{1/(1-\beta)} \quad \text{for} \quad x = \rho t^{\frac{m-\beta}{2(1-\beta)}}, \quad t \to 0+, \tag{2.16}$$

where the right-hand side of (2.16) (respectively, (2.15)) relates to the self-similar solution (2.9) (respectively, to its interface, as in (2.10)). If $m+\beta = 2$ we then have explicit values of ζ_* and $f_1(\rho)$ via (2.8), while in general we have lower and upper bounds via (2.11)–(2.14). If u_0 satisfies (1.3) with $\alpha = 2/(m-\beta)$, $C = C_*$, then the small-time behavior of the interface and the local solution depend on the terms smaller than $C_*(-x)^{2/(m-\beta)}$ in the expansion of u_0 as $x \to 0-$.

(3) Suppose that $b > 0$, $0 < \beta < 1$, $\alpha > 2/(m-\beta)$ (here again we describe the results for the case $m = 1$ as well). In this case the interface initially shrinks and

$$\eta(t) \sim -\ell_* t^{1/\alpha(1-\beta)} \quad \text{as} \quad t \to 0+, \tag{2.17}$$

where $\ell_* = C^{-1/\alpha}(b(1-\beta))^{1/\alpha(1-\beta)}$. For arbitrary $\ell > \ell_*$ we have

$$u(x,t) \sim \Big[C^{1-\beta}(-x)_+^{\alpha(1-\beta)} - b(1-\beta)t\Big]^{1/(1-\beta)} \quad \text{as} \quad t \to 0+ \tag{2.18}$$

along the curve $x = \eta_\ell(t) = -\ell t^{1/\alpha(1-\beta)}$. These results mean that the interface initially coincides with that of the solution

$$\bar{u}(x,t) = \Big[C^{1-\beta}(-x)_+^{\alpha(1-\beta)} - b(1-\beta)t\Big]_+^{1/(1-\beta)}$$

to the problem

$$\bar{u}_t + b\bar{u}^\beta = 0, \qquad \bar{u}(x,0) = C(-x)_+^\alpha.$$

Respective lower and upper estimations are given in section 4 (see (4.16) and (4.19) below).

(4) In this case the interface initially remains stationary. We divide the results into four different subcases (see Figure 1).

(4a) Let $\beta = 1$, $\alpha = 2/(m-1)$. This case reduces to the case $b = 0$ by a simple transformation (see section 3). If u_0 is defined by (1.4), then the unique solution [8, 20] to (1.1), (1.4) is

$$\begin{aligned} &u_C(x,t) = C(-x)_+^{2/(m-1)} \\ &\times \exp(-bt)\Big[1 - (C/\bar{C})^{m-1} b^{-1}\Big(1 - \exp(-b(m-1)t)\Big)\Big]^{1/1-m} \end{aligned} \tag{2.19}$$

for $x \in \mathbb{R},\ t \in [0,T)$, where

$$T = +\infty \quad \text{if} \quad b \geq (C/\bar{C})^{m-1},$$
$$T = \Big(b(1-m)\Big)^{-1} \ln\big[1 - b(\bar{C}/C)^{m-1}\big] \quad \text{if} \quad -\infty < b < (C/\bar{C})^{m-1},$$
$$\bar{C} = \Big[(m-1)^2 \Big/ \Big(2m(m+1)\Big)\Big]^{1/(m-1)}.$$

If u_0 satisfies (1.3), then lower and upper estimations are given by $u_{C\mp\varepsilon}$.

(4b) Let $\beta = 1,\ \alpha > 2/(m-1)$. Then for arbitrary sufficiently small $\varepsilon > 0$, there exist $x_\varepsilon < 0$ and $\delta_\varepsilon > 0$ such that

$$(C-\varepsilon)(-x)_+^\alpha \exp(-bt) \leq u(x,t) \leq (C+\varepsilon)(-x)_+^\alpha \exp(-bt)$$
$$\times \left[1 - \varepsilon\Big(b(m-1)\Big)^{-1}\Big(1 - \exp(-b(m-1)t)\Big)\right]^{1/(1-m)}, \quad x \geq x_\varepsilon,\ \ 0 \leq t \leq \delta_\varepsilon. \tag{2.20}$$

(4c) Let $1 < \beta < m,\ \alpha \geq 2/(m-\beta)$. Then for arbitrary sufficiently small $\varepsilon > 0$ there exist $x_\varepsilon < 0$ and $\delta_\varepsilon > 0$ such that

$$g_{-\varepsilon}(x,t) \leq u(x,t) \leq g_\varepsilon(x,t), \quad x \geq x_\varepsilon,\ \ 0 \leq t \leq \delta_\varepsilon, \tag{2.21}$$

where

$$g_\varepsilon(x,t) = \begin{cases} \big[(C+\varepsilon)^{1-\beta}|x|^{\alpha(1-\beta)} + b(\beta-1)(1-d_\varepsilon)t\big]^{1/(1-\beta)}, & x_\varepsilon \leq x < 0, \\ 0, & x \geq 0, \end{cases}$$

$$d_\varepsilon = \begin{cases} \varepsilon \operatorname{sign} b & \text{if } \alpha > 2/(m-\beta), \\ \left(\Big((C+\varepsilon)/C_*\Big)^{m-\beta} + \varepsilon\right) \operatorname{sign} b & \text{if } \alpha = 2/(m-\beta), \end{cases}$$

and the constant C_* is defined in (I(2)).

(4d) Let either $1 < \beta < m,\ 2/(m-1) \leq \alpha < 2/(m-\beta)$, or $\beta \geq m,\ \alpha \geq 2/(m-1)$. If $\alpha = 2/(m-1)$ then for arbitrary $\varepsilon > 0$ there exist $x_\varepsilon < 0$ and $\delta_\varepsilon > 0$ such that

$$\begin{aligned} &(C-\varepsilon)(-x)_+^{2/(m-1)}(1-\gamma_{-\varepsilon}t)^{1/(1-m)} \leq u \\ &\qquad \leq (C+\varepsilon)(-x)_+^{2/(m-1)}(1-\gamma_\varepsilon t)^{1/(1-m)}, \quad x \geq x_\varepsilon,\ 0 \leq t \leq \delta_\varepsilon, \end{aligned} \tag{2.22}$$

where

$$\gamma_\varepsilon = \Big[2m(m+1)(C+\varepsilon)^{m-1} \Big/ (m-1)\Big] + \varepsilon.$$

However, if $\alpha > 2/(m-1)$ then for arbitrary $\varepsilon > 0$ there exist $x_\varepsilon < 0$ and $\delta_\varepsilon > 0$ such that

$$(C-\varepsilon)(-x)_+^\alpha \leq u \leq (C+\varepsilon)(-x)_+^\alpha (1-\varepsilon t)^{1/(1-m)}, \quad x \geq x_\varepsilon,\ \ 0 \leq t \leq \delta_\varepsilon. \tag{2.23}$$

(II) $b = 0$

REMARK 2.1. *It should be noted that this case has been widely investigated earlier* [7, 8, 14, 15, 28, 30, 31, 36] *(see also the review article* [23]*). Nevertheless, using the*

same techniques as in the case $b \neq 0$, we derive some global estimations (see (2.27)*). The new element here is that we have constructed lower and upper solutions to the corresponding nonlinear ODE for the function $f(\xi)$ in* (2.25).

(1) Let $m > 1,\ 0 < \alpha < 2/(m-1)$. In this case the interface expands. First, assume that $u_0(x)$ is defined by (1.4). Then if $\alpha = 1/(m-1)$ we have an explicit solution to the problem (1.1), (1.4):

$$u(x,t) = C(\xi_* t - x)_+^{1/(m-1)}, \quad \xi_* = C^{m-1} m (m-1)^{-1}. \tag{2.24}$$

If $0 < \alpha < 2/(m-1)$, then the solution to (1.1), (1.4) has the self-similar form

$$u(x,t) = t^{\frac{\alpha}{2+\alpha(1-m)}} f(\xi), \qquad \xi = x t^{-\frac{1}{2+\alpha(1-m)}}, \tag{2.25}$$

$$\eta(t) \quad = \xi_* t^{\frac{1}{2+\alpha(1-m)}}, \qquad 0 \le t < +\infty, \tag{2.26}$$

where ξ_* and f satisfy (2.2), (2.4)–(2.6). Moreover, we have

$$C_4 t^{\frac{\alpha}{2+\alpha(1-m)}} \Big(\xi_3 - \xi\Big)_+^{\frac{1}{m-1}} \le u \le C_5 t^{\frac{\alpha}{2+\alpha(1-m)}} \Big(\xi_4 - \xi\Big)_+^{\frac{1}{m-1}}, \tag{2.27}$$
$$0 \le x < +\infty, \quad 0 < t < +\infty,$$

where ξ_3 (respectively, ξ_4) is defined by the right-hand side of (2.5), where we replace ξ_*'' with $C^{\frac{m-1}{2-\alpha(m-1)}} \xi_1$ (respectively, with $C^{\frac{m-1}{2-\alpha(m-1)}} \xi_2$) and

$$C_4 = C^{2/(2-\alpha(m-1))} A_0 \xi_3^{1/(1-m)}, \qquad C_5 = C^{2/(2-\alpha(m-1))} A_0 \xi_4^{1/(1-m)}.$$

In the particular case $\alpha = (m-1)^{-1}$, when an explicit solution is given by (2.24), we have $\xi_3 = \xi_4 = \xi_*$ and both lower and upper estimations in (2.27) lead to the explicit solution (2.24). In general, when $\alpha \neq (m-1)^{-1}$ the precise value ξ_* relates to the similarity ODE for $f(\xi)$ from (2.25), namely, $\xi_* = \sup\{\xi : f(\xi) > 0\}$. If u_0 satisfies (1.3) with $0 < \alpha < 2/(m-1)$, then (2.1) and (2.3) are valid. Lower and upper bounds for $f(\rho)$ follow from (2.27).

(2) Let $m > 1,\ \alpha = 2/(m-1)$. In this case the interface initially remains stationary. If u_0 is defined by (1.4), then the explicit solution [8, 20] to (1.1), (1.4) is

$$u_C(x,t) = C(-x)_+^{2/(m-1)} \left[1 - (C/\bar{C})^{m-1}(m-1)t\right]^{1/(1-m)}, \quad x \in \mathbb{R}, \quad 0 \le t < T, \tag{2.28}$$

where

$$T = (\bar{C}/C)^{m-1}(m-1)^{-1}$$

and the constant $\bar{C}$ is defined in (I(4)).
If u_0 satisfies (1.3) with $\alpha = 2/(m-1)$, then lower and upper estimations are given by $u_{C \mp \varepsilon}$.

(3) Let $m > 1, \alpha > 2/(m-1)$. In this case also the interface initially remains stationary and for arbitrary sufficiently small $\varepsilon > 0$ there exist $x_\varepsilon < 0$ and $\delta_\varepsilon > 0$ such that

$$(C-\varepsilon)(-x)_+^{\alpha} \le u \le (C+\varepsilon)(-x)_+^{\alpha}(1-\varepsilon t)^{1/(1-m)}, \quad x_\varepsilon \le x, \quad 0 \le t \le \delta_\varepsilon. \tag{2.29}$$

3. Preliminary results. The mathematical theory of nonlinear degenerate parabolic equations began with the paper [32]. The methods and the results of [32] have been developed in [16, 21, 26] for more general equations (including (1.1) as a particular case). Significant progress in high-dimensional problems has been made due to the papers [5, 11, 12]. For a general study we refer to the survey article [23]. Boundary value problems for (1.1) in noncylindrical domains with nonsmooth boundaries have recently been investigated in [3].

Throughout this paper we shall follow the definitions of generalized solutions and of supersolutions (or subsolutions) of initial or initial boundary value problems to (1.1) given in [3]. Comparison theorems from [3] for the solutions of Cauchy–Dirichlet and Dirichlet problems in irregular domains will be extensively used in the proof of the results of section 2.

REMARK 3.1. *It should be noted that in the paper* [3] *it was supposed that solutions (respectively, supersolutions and subsolutions) of the Cauchy–Dirichlet problem are bounded on every bounded time interval. This assumption had no importance in the proof of existence theorems, since the constructed limit solutions possess this property, but it has been used in the proof of uniqueness theorems in* [3]. *However, by using a slight technical modification of the proofs from* [3] *in the case of* $b \geq 0$, *the uniqueness and comparison assertions of Theorems* 2.2 *and* 2.4, *and also Lemma* 2.1 *from* [3], *may be proved without this assumption.*

Suppose that $b \geq 0$ and that u_0 may have unbounded growth as $|x| \to +\infty$. It is well known that in this case some necessary and sufficient conditions must be imposed on the growth rate for existence, uniqueness, and comparison results in the CP (1.1), (1.2). For the slow diffusion equation ((1.1) with $b = 0,\ m > 1$) the optimal growth condition is known due to [9, 20]. In particular, if initial data may be majorized by power law function (1.4), then there exists a unique solution (with $T = +\infty$) and a comparison principle is valid if $0 < \alpha < 2/(m-1)$. If $\alpha = 2/(m-1)$, then existence, uniqueness, and comparison results are valid only locally in time. For instance, from [8, 20] it follows that the unique explicit solution to (1.1), (1.4) with $b = 0,\ \alpha = 2/(m-1), T = (\bar{C}/C)^{m-1}(m-1)^{-1}$ is $u_C(x,t)$ from (2.28).

If the function $u(x,t)$ is a solution to (1.1) with $b = 0$, then the function

$$\bar{u}(x,t) = \exp(-bt)\ u(x, (b(1-m))^{-1}(\exp(b(1-m)t) - 1))$$

is a solution to (1.1) with $b \neq 0,\ \beta = 1$. Hence, from the above mentioned result it follows that the unique solution to CP (1.1), (1.4) with $m > 1,\ b \neq 0,\ \beta = 1,\ \alpha = 2/(m-1)$ is the function $\bar{u}_C(x,t)$ from (2.19).

Necessary and sufficient conditions on the growth rate at infinity for existence, uniqueness, and comparison results for the CP (1.1), (1.2) with $b > 0,\ m > 1,\ \beta > 0$ have been investigated in [22, 24, 6, 1]. We are not interested in describing an optimal result; for our purposes it is enough to mention that if u_0 may be majorized by the function (1.4) with α satisfying $0 < \alpha < 2/(m-1)$, then the CP (1.1), (1.2) with $b > 0,\ m > 1,\ \beta > 0,\ T = +\infty$ has a unique solution and for this class of initial data a comparison principle is valid [24, 6].

In the next four lemmas we establish some preliminary estimations of the solution to CP. The proof of these estimations is based on scale of variables.

LEMMA 3.1. *If* $b = 0$ *and* $m > 1,\ 0 < \alpha < 2/(m-1)$, *then the solution* u *of the CP* (1.1), (1.4) *has a self-similar form* (2.25), *where the self-similarity function* f *satisfies* (2.4). *If* u_0 *satisfies* (1.3), *then the solution to CP* (1.1), (1.2) *satisfies* (2.1)–(2.3).

LEMMA 3.2. *Let u be a solution to the CP* (1.1), (1.2) *and u_0 satisfy* (1.3). *Let one of the following conditions be valid:*

(a) $b > 0, \quad 0 < \beta < 1 < m, \quad 0 < \alpha < 2/(m-\beta)$,

(b) $b \neq 0, \quad \beta \geq 1, \quad m > 1, \quad 0 < \alpha < 2/(m-1)$.

Then u satisfies (2.3).

LEMMA 3.3. *Let u be a solution to CP* (1.1), (1.4) *with* $b > 0,\ 0 < \beta < 1,\ m \geq 1, \alpha = 2/(m-\beta)$. *Then the solution u has the self-similar form* (2.9). *If* $C > C_*$ *then $f_1(0) = A_1$, where A_1 is a positive number depending on m, β, C, and b. If u_0 satisfies* (1.3) *with* $\alpha = 2/(m-\beta),\ C > C_*,$ *then u satisfies*

$$u(0,t) \sim A_1\, t^{1/(1-\beta)} \quad \text{as} \quad t \to 0+. \tag{3.1}$$

LEMMA 3.4. *Let u be a solution to the CP* (1.1)–(1.3) *with* $b > 0,\ 0 < \beta < 1,\ \alpha > 2/(m-\beta)$. *Then for arbitrary $\ell > \ell_*$ (see* (2.17)*) the asymptotic formula* (2.18) *is valid with $x = \eta_\ell(t) = -\ell t^{1/\alpha(1-\beta)}$.*

Proof of Lemma 3.1. The self-similar form (2.25) of the solution to (1.1), (1.4) is a well-known result (see Remark 2.1 in section 2). If we consider a function

$$u_\kappa(x,t) = \kappa u\left(\kappa^{-1/\alpha}x,\ \kappa^{(\alpha(m-1)-2)/\alpha}t\right), \quad \kappa > 0, \tag{3.2}$$

it may easily be checked that this satisfies (1.1), (1.4). Since under the condition of the lemma there exists a unique global solution to (1.1), (1.4), we have

$$u(x,t) = \kappa u\left(\kappa^{-1/\alpha}x,\ \kappa^{(\alpha(m-1)-2)/\alpha}t\right), \quad \kappa > 0. \tag{3.3}$$

If we choose $\kappa = t^{\alpha/(2-\alpha(m-1))}$, then (3.3) implies (2.25) with $f(\xi) = u(\xi,1)$. It is a well-known fact (see Remark 2.1) that f is a unique nonnegative and continuous solution of the relevant boundary value problem for the similarity ODE and there exists an $\xi_* > 0$ such that f is positive and smooth for $\xi < \xi_*$ and $f = 0$ for $\xi \geq \xi_*$. Thus, (2.26) is valid. Now, to find the dependence of f on C we can again use scaling as in [36]. Namely, let ω be a solution of the CP (1.1), (1.4) with $C = 1$. Then it may be easily checked that for arbitrary $\kappa > 0$

$$u(x,t) = \kappa\omega\left(C^{1/\alpha}\kappa^{-1/\alpha}x, \quad C^{2/\alpha}\kappa^{(\alpha(m-1)-2)/\alpha}t\right).$$

By choosing $\kappa = \left(C^{2/\alpha}t\right)^{\alpha/(2-\alpha(m-1))}$ we then have

$$u(x,t) = C^{\frac{2}{2-\alpha(m-1)}}\omega\left(C^{\frac{m-1}{\alpha(m-1)-2}}\xi, 1\right)t^{\alpha/(2-\alpha(m-1))}. \tag{3.4}$$

From (3.4) and (2.25), (2.4) and (2.2) follow.

Now suppose that u_0 satisfies (1.3). Then for arbitrary sufficiently small $\varepsilon > 0$ there exists an $x_\varepsilon < 0$ such that

$$(C-\varepsilon/2)(-x)_+^\alpha \leq u_0(x) \leq (C+\varepsilon/2)(-x)_+^\alpha, \qquad x \geq x_\varepsilon. \tag{3.5}$$

Let $u_\varepsilon(x,t)$ (respectively, $u_{-\varepsilon}(x,t)$) be a solution to the CP (1.1), (1.2) with initial data $(C+\varepsilon)(-x)_+^\alpha$ (respectively, $(C-\varepsilon)(-x)_+^\alpha$). Since the solution to the CP (1.1), (1.2) is continuous there exists a number $\delta = \delta(\varepsilon) > 0$ such that

$$u_\varepsilon(x_\varepsilon,t) \geq u(x_\varepsilon,t), \quad u_{-\varepsilon}(x_\varepsilon,t) \leq u(x_\varepsilon,t) \quad \text{for} \quad 0 \leq t \leq \delta. \tag{3.6}$$

From (3.5), (3.6), and a comparison principle (e.g., Theorem 2.4 of [3]), it follows that

$$u_{-\varepsilon} \le u \le u_{\varepsilon} \quad \text{for} \quad x \ge x_{\varepsilon}, \quad 0 \le t \le \delta. \tag{3.7}$$

Obviously

$$u_{\pm\varepsilon}(\xi_\rho(t), t) = f(\rho; C \pm \varepsilon) t^{\alpha/(2-\alpha(m-1))}, \quad t \ge 0. \tag{3.8}$$

(Furthermore, we denote the right-hand side of (2.4a) by $f(\rho, C)$.) Now taking $x = \xi_\rho(t)$ in (3.7), after multiplying to $t^{-\alpha/(2-\alpha(m-1))}$ and passing to the limit, first as $t \to 0$ and then as $\varepsilon \to 0$, we can easily derive (2.3). Similarly, from (3.7), (2.26), and (2.2), (2.1) easily follows. The lemma is proved.

Proof of Lemma 3.2. As in the proof of Lemma 3.1, (3.5) and (3.6) follow from (1.3). Let the conditions of one of the cases (a) or (b) with $b > 0$ be valid. Then from results mentioned earlier it follows that the existence, uniqueness, and comparison results of the CP (1.1), (1.2) with $u_0 = (C \pm \varepsilon)(-x)_+^\alpha$, $T = +\infty$ hold. As before, from (3.5) and (3.6), (3.7) follows. Now if we take

$$u_\kappa^{\pm\varepsilon}(x,t) = \kappa\, u_{\pm\varepsilon}\left(\kappa^{-1/\alpha} x,\ \kappa^{(\alpha(m-1)-2)/\alpha} t\right), \quad \kappa > 0, \tag{3.9}$$

then $u_\kappa^{\pm\varepsilon}(x,t)$ satisfies the following problem:

$$u_t - (u^m)_{xx} + b\kappa^{(\alpha(m-\beta)-2)/\alpha} u^\beta = 0, \quad x \in \mathbb{R}, \quad t > 0, \tag{3.10a}$$

$$u(x,0) = (C \pm \varepsilon)(-x)_+^\alpha, \quad x \in \mathbb{R}. \tag{3.10b}$$

There exists a unique solution to CP (3.10), which obeys a comparison principle also. Since $\alpha(m-\beta) - 2 < 0$, by using a comparison principle it may easily be proved that

$$\lim_{\kappa \to +\infty} u_\kappa^{\pm\varepsilon}(x,t) = v_{\pm\varepsilon}(x,t), \quad x \in \mathbb{R}, \quad t \ge 0, \tag{3.11}$$

where $v_{\pm\varepsilon}$ is a solution to the CP (1.1), (1.2) with $b = 0$, $u_0 = (C \pm \varepsilon)(-x)_+^\alpha$, $T = +\infty$. Hence, $v_{\pm\varepsilon}$ satisfies (3.8). If we now take $x = \xi_\rho(t)$, where ρ is an arbitrary fixed number satisfying $\rho < \xi_*$, then from (3.11) it follows that

$$\lim_{\kappa \to +\infty} \kappa u_{\pm\varepsilon}\left(\kappa^{-1/\alpha}\xi_\rho(t),\ \kappa^{(\alpha(m-1)-2)/\alpha} t\right) = f(\rho; C \pm \varepsilon) t^{\alpha/(2-\alpha(m-1))}, \quad t > 0. \tag{3.12}$$

If we take $\tau = \kappa^{(\alpha(m-1)-2)/\alpha} t$, then (3.12) implies

$$u_{\pm\varepsilon}(\xi_\rho(\tau), \tau) \sim f(\rho; C \pm \varepsilon)\tau^{\alpha/(2-\alpha(m-1))} \ \text{ as } \ \tau \to 0+. \tag{3.13}$$

As before, (2.3) easily follows from (3.7), (3.13).

Now consider the case (b) with $b < 0$. Suppose that $u_{\pm\varepsilon}$ is a solution of the Dirichlet problem

$$u_t - (u^m)_{xx} + bu^\beta = 0, \qquad |x| < |x_\varepsilon|, \quad 0 < t \le \delta, \tag{3.14a}$$

$$u(x,0) = (C \pm \varepsilon)(-x)_+^\alpha, \qquad |x| \le |x_\varepsilon|, \tag{3.14b}$$

$$u(x_\varepsilon, t) = (C \pm \varepsilon)(-x_\varepsilon)^\alpha, \qquad u(-x_\varepsilon, t) = 0, \quad 0 \le t \le \delta. \tag{3.14c}$$

The function $u_\kappa^{\pm\varepsilon}$, defined as in (3.9), satisfies the Dirichlet problem

$$u_t - (u^m)_{xx} + b\kappa^{(\alpha(m-\beta)-2)/\alpha}u^\beta = 0 \quad \text{in} \quad D_\varepsilon^\kappa,$$

(3.15a)

$$u\left(\kappa^{1/\alpha}x_\varepsilon, t\right) = \kappa(C \pm \varepsilon)(-x_\varepsilon)^\alpha, \quad u\left(-\kappa^{1/\alpha}x_\varepsilon, t\right) = 0, \quad 0 \le t \le \kappa^{(2-\alpha(m-1))/\alpha}\delta,$$

(3.15b)

$$u(x,0) = (C \pm \varepsilon)(-x)_+^\alpha, \qquad |x| \le \kappa^{1/\alpha}|x_\varepsilon|,$$

(3.15c)

where

$$D_\varepsilon^\kappa = \left\{(x,t):\ |x| < \kappa^{1/\alpha}|x_\varepsilon|, \qquad 0 < t \le \kappa^{(2-\alpha(m-1))/\alpha}\delta\right\}.$$

From Theorem 3.1 of [3] it follows that there exists a number $\delta > 0$ (which does not depend on κ) such that both (3.14) and (3.15) have a unique solution.

In view of finite speed of propagation a $\delta = \delta(\varepsilon) > 0$ may be chosen such that

$$u(-x_\varepsilon, t) = 0, \qquad 0 \le t \le \delta. \tag{3.16}$$

Applying the Comparison Theorem 3.4 of [3], from (3.5), (3.6), and (3.16), (3.7) follows for $|x| \le |x_\varepsilon|$, $0 \le t \le \delta$.

The next step consists in the proof of the convergence of the sequences $\{u_\kappa^{\pm\varepsilon}\}$ as $\kappa \to +\infty$. Consider a function

$$g(x,t) = (C+1)(1+x^2)^{\alpha/2}(1-\nu t)^{1/(1-m)}, \quad x \in \mathbb{R}, \quad 0 \le t \le t_0 = \nu^{-1}/2,$$

where

$$\nu = h_* + 1, \quad h_* = h_*(\alpha; m) = \max_{x\in\mathbb{R}} h(x),$$
$$h(x) = (m-1)(C+1)^{m-1}\alpha m(1+x^2)^{\alpha(m-1)/2-2}\left(1 + (\alpha m - 1)x^2\right).$$

Then we have

$$\begin{aligned} \mathrm{L}_\kappa g \equiv g_t - (g^m)_{xx} + b\kappa^{(\alpha(m-\beta)-2)/\alpha}g^\beta \\ = (C+1)(m-1)^{-1}(1+x^2)^{\alpha/2}(1-\nu t)^{m/(1-m)}\mathrm{S} \quad \text{in} \quad D_\varepsilon^\kappa, \\ \mathrm{S} = \nu - h(x) + b(m-1)(C+1)^{\beta-1}\kappa^{(\alpha(m-\beta)-2)/\alpha} \\ \times (1+x^2)^{\alpha(\beta-1)/2}(1-\nu t)^{(\beta-m)/(1-m)}, \end{aligned}$$

and hence

$$\mathrm{S} \ge 1 + R \ \text{ in } \ D_{0\varepsilon}^\kappa = D_\varepsilon^\kappa \cap \left\{(x,t):\ 0 < t \le t_0\right\}, \tag{3.17}$$

where

$$R = O(\kappa^{m-1-2/\alpha}) \ \text{ uniformly for } \ (x,t) \in D_{0\varepsilon}^\kappa \quad \text{as} \quad \kappa \to +\infty.$$

Moreover, we have for $0 < \varepsilon \ll 1$

$$g(x,0) \ge u_\kappa^{\pm\varepsilon}(x,0) \qquad \text{for} \ \ |x| \le \kappa^{1/\alpha}|x_\varepsilon|, \tag{3.18a}$$

$$g(\pm\kappa^{1/\alpha}x_\varepsilon, t) \ge u_\kappa^{\pm\varepsilon}\left(\pm\kappa^{1/\alpha}x_\varepsilon, t\right) \quad \text{for} \ \ 0 \le t \le t_0. \tag{3.18b}$$

Hence, there exists a number $\kappa_0 = \kappa_0(\alpha; m)$ such that for $\kappa \geq \kappa_0$ the Comparison Theorem 3.4 of [3] implies

$$0 \leq u_\kappa^{\pm\varepsilon}(x,t) \leq g(x,t) \quad \text{in} \quad \bar{D}_{0\varepsilon}^\kappa. \tag{3.19}$$

Let G be an arbitrary fixed compact subset of

$$P = \Big\{(x,t):\ x \in \mathbb{R}, \quad 0 < t \leq t_0\Big\}.$$

We take κ_0 so large that $G \subset D_{0\varepsilon}^\kappa$ for $\kappa \geq \kappa_0$. From (3.19) it follows that the sequences $\{u_\kappa^{\pm\varepsilon}\}$, $\kappa \geq \kappa_0$, are uniformly bounded in G. As in [3], it may be proved that they are uniformly Hölder continuous in G and that there exist functions $v_{\pm\varepsilon}$ such that for some subsequence κ'

$$\lim_{\kappa' \to +\infty} u_{\kappa'}^{\pm\varepsilon}(x,t) = v_{\pm\varepsilon}(x,t), \quad (x,t) \in P. \tag{3.20}$$

It may easily be checked that $v_{\pm\varepsilon}$ is a solution to the CP (1.1), (1.2) with $b = 0$, $T = t_0, u_0 = (C \pm \varepsilon)(-x)_+^\alpha$. As before, from (3.8), (3.12), (3.13), and (3.7), the required estimation (2.3) follows. The lemma is proved.

The first assertion of Lemma 3.3 has been proved in [6] for the case $m > 1$. If u_0 satisfies (1.3), the estimation (3.1) may be proved exactly as estimation (2.3) was proved in Lemma 3.1. Finally, the case $m = 1$ may be considered similarly.

Proof of Lemma 3.4. As before, (3.5) and (3.6) follow from (1.3). Suppose that $u_{\pm\varepsilon}$ is a solution of the problem

$$\begin{aligned}
&v_t - (v^m)_{xx} + bv^\beta = 0, && |x| < |x_\varepsilon|, \quad 0 < t \leq \delta,\\
&v(x,0) = (C \pm \varepsilon)(-x)_+^\alpha, && |x| \leq |x_\varepsilon|,\\
&v(x_\varepsilon, t) = (C \pm \varepsilon)(-x_\varepsilon)^\alpha, \quad v(-x_\varepsilon, t) = u(-x_\varepsilon, t), && 0 \leq t \leq \delta.
\end{aligned}$$

Applying a comparison principle (e.g., Theorem 3.4 of [3]), from (3.5) and (3.6), (3.7) follows for $|x| \leq |x_\varepsilon|$, $0 \leq t \leq \delta$. Now if we take

$$u_\kappa^{\pm\varepsilon}(x,t) = \kappa u_{\pm\varepsilon}\left(\kappa^{-1/\alpha}x,\ \kappa^{\beta-1}t\right), \quad \kappa > 0,$$

then $u_\kappa^{\pm\varepsilon}$ satisfies the Dirichlet problem

$$\begin{aligned}
&v_t - \kappa^{\frac{2-\alpha(m-\beta)}{\alpha}}(v^m)_{xx} + bv^\beta = 0 && \text{in} \quad E_\varepsilon^\kappa,\\
&v(x,0) = (C \pm \varepsilon)(-x)_+^\alpha, && |x| \leq \kappa^{1/\alpha}|x_\varepsilon|,\\
&v\left(\kappa^{1/\alpha}x_\varepsilon, t\right) = \kappa(C \pm \varepsilon)(-x_\varepsilon)^\alpha, && v\left(-\kappa^{1/\alpha}x_\varepsilon, t\right) = \kappa u\left(-x_\varepsilon, \kappa^{\beta-1}t\right),\\
&\quad 0 \leq t \leq \kappa^{1-\beta}\delta,
\end{aligned}$$

where

$$E_\varepsilon^\kappa = \Big\{(x,t):\ |x| < \kappa^{1/\alpha}|x_\varepsilon|, \quad 0 < t \leq \kappa^{1-\beta}\delta\Big\}.$$

The next step consists in proving the convergence of the sequences $\{u_\kappa^{\pm\varepsilon}\}$ as $\kappa \to +\infty$. Considering the function $g(x,t) = (C+1)(1+x^2)^{\alpha/2}\exp t$, we have

$$\begin{aligned}
\tilde{\mathrm{L}}_\kappa g \equiv g_t - \kappa^{\frac{2-\alpha(m-\beta)}{\alpha}}(g^m)_{xx} + bg^\beta \geq g\Big[1 - (C+1)^{m-1}\alpha m(1+x^2)^{\frac{\alpha(m-1)}{2}-2}\\
\times\Big(1 + (\alpha m - 1)x^2\Big)\exp\Big((m+1)t\Big)\kappa^{\frac{2-\alpha(m-\beta)}{\alpha}}\Big] \text{ in } E_\varepsilon^\kappa.
\end{aligned} \tag{3.21}$$

Let $t_0 > 0$ be fixed and let $E^\kappa_{0_\varepsilon} = E^\kappa_\varepsilon \cap \{(x,t) : \ 0 < t \le t_0\}$. Then from (3.21) it follows that

$$\tilde{\mathrm{L}}_\kappa g \ge g(1+R) \quad \text{in} \ \ E^\kappa_{0_\varepsilon},$$

where

$$R = O(\kappa^\theta) \text{ uniformly for } (x,t) \in E^\kappa_{0_\varepsilon} \ \text{ as } \ \kappa \to +\infty,$$

$$\theta \ = \Big(2 - \alpha(m-\beta)\Big)\Big/\alpha \quad \text{if} \ \ \alpha < 2/(m-1),$$

$$\theta \ = \beta - 1 \qquad \text{if} \ \ \alpha \ge 2/(m-1).$$

Moreover, we have for $0 < \varepsilon \ll 1$ that

$$g(x,0) \ge u^{\pm\varepsilon}_\kappa(x,0) \quad \text{for} \ \ |x| \le \kappa^{1/\alpha}|x_\varepsilon|.$$

Since

$$u^{\pm\varepsilon}_\kappa\left(-\kappa^{1/\alpha}x_\varepsilon, t\right) = o(\kappa) \quad \text{for} \ \ 0 \le t \le t_0 \quad \text{as } \kappa \to +\infty,$$

we also have

$$g\left(\pm\kappa^{1/\alpha}x_\varepsilon, t\right) \ge u^{\pm\varepsilon}_\kappa\left(\pm\kappa^{1/\alpha}x_\varepsilon, t\right) \quad \text{for} \ \ 0 \le t \le t_0$$

if κ is chosen large enough. Hence, as in the proof of Lemma 3.2, if κ is large enough, a comparison principle (e.g., Theorem 3.4 of [3]) implies (3.19) in $\bar{E}^\kappa_{0_\varepsilon}$, where the respective functions $u^{\pm\varepsilon}_\kappa$ and g apply in the context of this proof. As in [3], it may then be proved that the sequences $\{u^{\pm\varepsilon}_\kappa\}$ are uniformly Hölder continuous on compact subsets of P. Thus there exist functions $v_{\pm\varepsilon}$ such that for some subsequence κ', (3.20) is valid. It may easily be shown that the limit functions $v_{\pm\varepsilon}$ are solutions to the problem

$$v_t + bv^\beta = 0, \quad x \in \mathbb{R}, \quad 0 < t \le t_0; \quad v(x,0) = (C \pm \varepsilon)(-x)^\alpha_+, \quad x \in \mathbb{R},$$

i.e.,

$$v_{\pm\varepsilon}(x,t) = \Big[(C \pm \varepsilon)^{1-\beta}(-x)^{\alpha(1-\beta)}_+ - b(1-\beta)t\Big]^{\frac{1}{1-\beta}}_+ .$$

Let $\ell > \ell_*$ be an arbitrary number and $\varepsilon > 0$ be chosen such that

$$(C-\varepsilon)^{1-\beta}\ell^{\alpha(1-\beta)} > b(1-\beta).$$

If we now take $x = \eta_\ell(t)$ and $\tau = \kappa^{\beta-1}t$, it follows from (3.20) that

$$u_{\pm\varepsilon}(\eta_\ell(\tau),\tau) \sim \Big[(C \pm \varepsilon)^{1-\beta}\ell^{\alpha(1-\beta)} - b(1-\beta)\Big]^{\frac{1}{1-\beta}} \tau^{\frac{1}{1-\beta}} \ \text{ as } \ \tau \to 0+ . \tag{3.22}$$

From (3.7) and (3.22), in view of the arbitrariness of $\varepsilon > 0$, the desired formula (2.18) follows easily. The lemma is proved.

REMARK 3.2. *Lemma* 3.4 *is true also if* $\beta < m \le 1$, *the proof completely coinciding with the one given. We just need to mention that* $\theta = (2 - \alpha(m-\beta))/\alpha$ *if* $\beta < m \le 1$.

4. Proofs of the main results. In this section we prove the main results described in section 2.

(I) $b \neq 0$ and $m > 1$.

(1) Suppose that $\alpha < 2/(m - \min\{1, \beta\})$ The formula (2.3) follows from Lemma 3.2. In view of the arbitrariness of ρ, it implies

$$\liminf_{t\to 0+} \eta(t) t^{1/(\alpha(m-1)-2)} \geq \xi_*. \tag{4.1}$$

Take an arbitrary sufficiently small number $\varepsilon > 0$. Let u_ε be a solution of the CP (1.1), (1.4) with $b = 0$ and with C replaced by $C + \varepsilon$. As before, the second inequality of (3.5) and the first inequality of (3.6) follow from (1.3). Suppose that $b > 0$. In this case, u_ε is a supersolution of (1.1). From (3.5), (3.6), and a comparison principle, the second inequality of (3.7) follows. By Lemma 3.1 we then have

$$\eta(t) \leq (C + \varepsilon)^{\frac{m-1}{2-\alpha(m-1)}} \xi'_* \, t^{1/(2-\alpha(m-1))}, \quad 0 \leq t \leq \delta,$$

and hence

$$\limsup_{t\to 0+} \eta(t) t^{1/(\alpha(m-1)-2)} \leq \xi_*. \tag{4.2}$$

Now suppose that $b < 0$ and $\beta \geq 1$. The function

$$\bar{u}_\varepsilon(x,t) = \exp(-bt) u_\varepsilon \left(x, \Big(b(1-m)\Big)^{-1} \Big[\exp\Big(b(1-m)t\Big) - 1\Big] \right)$$

is a solution to the CP (1.1),(1.4) with $\beta = 1$ and with C replaced by $C + \varepsilon$. As before, from (1.3) the first inequality of (3.6) follows, where we replace u_ε with $\bar{u}_\varepsilon$. Then we make $|x_\varepsilon|$ and δ so small that

$$\bar{u}_\varepsilon < 1 \ \text{ in } \ B = \Big\{(x,t) : \ x \geq x_\varepsilon, \ \ 0 < t \leq \delta\Big\}.$$

Obviously, $\bar{u}_\varepsilon$ is a supersolution of (1.1) in B. From (3.5), (3.6), and a comparison principle, the second inequality of (3.7), with u_ε replaced by $\bar{u}_\varepsilon$, follows. Thus we have

$$\eta(t) \leq (C + \varepsilon)^{\frac{m-1}{2-\alpha(m-1)}} \xi'_* \left\{ \Big(b(1-m)\Big)^{-1} \Big[\exp\Big(b(1-m)t\Big) - 1\Big] \right\}^{1/(2-\alpha(m-1))}, \quad 0 \leq t \leq \delta,$$

which again implies (4.2). From (4.1) and (4.2), (2.1) follows. Finally, (2.5), (2.6), (2.7) follow from (2.27), which we prove later in this section.

(2) $b > 0, \ 0 < \beta < 1, \ m \geq 1, \ \alpha = 2/(m - \beta)$.

First, assume that u_0 is defined by (1.4). As mentioned earlier in section 3, the problem (1.1), (1.4) has a unique global solution and for this class of initial data a comparison principle is valid.

If $m + \beta = 2$ it may be easily checked that the explicit solution to (1.1), (1.4) is given by (2.8).

Let $m + \beta \neq 2$. The self-similar form (2.9) follows from Lemma 3.3. Let $C > C_*$. Consider a function

$$g(x,t) = t^{1/(1-\beta)} f_1(\zeta), \qquad \zeta = x \, t^{-\frac{m-\beta}{2(1-\beta)}}. \tag{4.3}$$

We then have

$$\mathrm{L}g = t^{\beta/(1-\beta)}\mathcal{L}^0 f_1, \tag{4.4a}$$

$$\mathcal{L}^0 f_1 = \frac{1}{1-\beta} f_1 - (f_1^m)'' - \frac{m-\beta}{2(1-\beta)}\zeta f_1' + b f_1^\beta. \tag{4.4b}$$

As a function f_1 we take

$$f_1(\zeta) = C_0(\zeta_0 - \zeta)_+^{\gamma_0}, \qquad 0 < \zeta < +\infty,$$

where $C_0,\ \zeta_0,\ \gamma_0$ are some positive constants. Taking $\gamma_0 = 2/(m-\beta)$, from (4.4b) we have

$$\mathcal{L}^0 f_1 = bC_0^\beta(\zeta_0-\zeta)^{\frac{2\beta}{m-\beta}}\left\{1 - \left(\frac{C_0}{C_*}\right)^{m-\beta} + \frac{C_0^{1-\beta}}{b(1-\beta)}\zeta_0(\zeta_0-\zeta)^{\frac{2-m-\beta}{m-\beta}}\right\}. \tag{4.5}$$

To prove an upper estimation we take $C_0 = C_2,\ \zeta_0 = \zeta_2$ (see Appendix). If $m+\beta > 2$, then we have

$$\mathcal{L}^0 f_1 \geq bC_2^\beta(\zeta_2-\zeta)^{\frac{2\beta}{m-\beta}}\left\{1 - \left(\frac{C_2}{C_*}\right)^{m-\beta} + \frac{C_2^{1-\beta}}{b(1-\beta)}\zeta_2^{\frac{2(1-\beta)}{m-\beta}}\right\} = 0 \quad \text{for } 0 \leq \zeta \leq \zeta_2,$$

while if $1 \leq m < 2-\beta$, we have

$$\mathcal{L}^0 f_1 \geq bC_2^\beta(\zeta_2-\zeta)^{\frac{2\beta}{m-\beta}}\left\{1 - \left(\frac{C_2}{C_*}\right)^{m-\beta}\right\} = 0 \quad \text{for } 0 \leq \zeta \leq \zeta_2.$$

From (4.4a) it follows that

$$\mathrm{L}g \geq 0 \quad \text{for } 0 < x < \zeta_2\, t^{\frac{m-\beta}{2(1-\beta)}}, \quad 0 < t < +\infty, \tag{4.6a}$$

$$\mathrm{L}g = 0 \quad \text{for } x > \zeta_2\, t^{\frac{m-\beta}{2(1-\beta)}}, \qquad 0 < t < +\infty. \tag{4.6b}$$

Lemma 2.1 of [3] implies that g is a supersolution of (1.1) in $\{(x,t):\ x>0,\ t>0\}$. Since

$$g(x,0) = u(x,0) = 0 \qquad \text{for } 0 \leq x < +\infty, \tag{4.7a}$$

$$g(0,t) = u(0,t) \qquad \text{for } 0 \leq t < +\infty, \tag{4.7b}$$

from Comparison Theorem 2.4 of [3], the right-hand side of (2.11) follows.
If $1 \leq m < 2-\beta$ then to prove the lower estimation we take $C_0 = C_1, \zeta_0 = \zeta_1,\ \gamma_0 = 2/(m-\beta)$. Then from (4.5) we derive

$$\mathcal{L}^0 f_1 \leq bC_1^\beta(\zeta_1-\zeta)^{\frac{2\beta}{m-\beta}}\left\{1 - \left(\frac{C_1}{C_*}\right)^{m-\beta} + \frac{C_1^{1-\beta}}{b(1-\beta)}\zeta_1^{\frac{2(1-\beta)}{m-\beta}}\right\} = 0 \quad \text{for } 0 \leq \zeta \leq \zeta_1,$$

and from (4.4a) it follows that

$$\mathrm{L}g \leq 0 \quad \text{for } 0 < x < \zeta_1\, t^{\frac{m-\beta}{2(1-\beta)}}, \quad 0 < t < +\infty, \tag{4.8a}$$

$$\mathrm{L}g = 0 \quad \text{for } x > \zeta_1\, t^{\frac{m-\beta}{2(1-\beta)}}, \qquad 0 < t < +\infty. \tag{4.8b}$$

As before, from Lemma 2.1 of [3], (4.7), and Comparison Theorem 2.4 of [3], the left-hand side of (2.11) follows.
If $m+\beta>2$, then to prove the lower estimation we take $C_0=C_1,\ \zeta_0=\zeta_1, \gamma_0=1/(m-1)$. Then from (4.4b) we have

$$\begin{aligned}
&\mathcal{L}^0 f_1 = C_1(1-\beta)^{-1}(\zeta_1-\zeta)^{\frac{2-m}{m-1}}\\
&\times\left\{\zeta_1-\frac{m+\beta-2}{2(m-1)}\zeta-C_1^{m-1}\frac{m(1-\beta)}{(m-1)^2}+b(1-\beta)C_1^{\beta-1}(\zeta_1-\zeta)^{\frac{m+\beta-2}{m-1}}\right\}\\
&\le C_1(1-\beta)^{-1}(\zeta_1-\zeta)^{\frac{2-m}{m-1}}\\
&\times\left\{\zeta_1-C_1^{m-1}m(1-\beta)(m-1)^{-2}+b(1-\beta)C_1^{\beta-1}\zeta_1^{\frac{m+\beta-2}{m-1}}\right\}=0 \qquad \text{for } 0<\zeta<\zeta_1,
\end{aligned}$$

which again implies (4.8). As before, from Lemma 2.1 of [3], (4.7), and Comparison Theorem 2.4 of [3], the left-hand side of (2.11) follows.

By applying the same analysis it may easily be checked that the alternative upper estimation is valid if $C_0=\bar{C}_2,\ \zeta_0=\bar{\zeta}_2,\ \gamma_0=1/(m-1)$.
Let $m+\beta>2$ and $0<C<C_*$. Consider a function

$$g(x,t)=\left[C^{1-\beta}(-x)_+^{\frac{2(1-\beta)}{m-\beta}}-b(1-\beta)(1-\gamma)t\right]_+^{\frac{1}{1-\beta}}, \qquad x\in\mathbb{R},\quad t>0,$$

where γ is a constant such that $\gamma\in[0,1)$. Let us estimate $\mathrm{L}g$ in

$$\begin{aligned}
M &= \Big\{(x,t):\ -\infty<x<\mu_\gamma(t),\quad t>0\Big\},\\
\mu_\gamma(t) &= -\Big[b(1-\beta)(1-\gamma)C^{\beta-1}t\Big]^{\frac{m-\beta}{2(1-\beta)}}.
\end{aligned}$$

We have

$$\mathrm{L}g=bg^\beta \mathrm{S},$$

$$\begin{aligned}
\mathrm{S}=\gamma&-2mb^{-1}(m-\beta)^{-2}(2-m-\beta)C^{m-\beta}\left[1-\frac{b(1-\beta)(1-\gamma)t}{C^{1-\beta}(-x)^{\frac{2(1-\beta)}{m-\beta}}}\right]^{\frac{m-1}{1-\beta}}\\
&-4mb^{-1}(m-\beta)^{-2}(m+\beta-1)C^{m-\beta}\left[1-\frac{b(1-\beta)(1-\gamma)t}{C^{1-\beta}(-x)^{\frac{2(1-\beta)}{m-\beta}}}\right]^{\frac{m+\beta-2}{1-\beta}}. \qquad \text{(4.9a)}
\end{aligned}$$

Hence

$$\mathrm{S}\Big|_{t=0}=\gamma-\left(\frac{C}{C_*}\right)^{m-\beta}, \qquad \mathrm{S}\Big|_{x=\mu_\gamma(t)}=\gamma. \qquad \text{(4.9b)}$$

Moreover

$$\begin{aligned}
&\mathrm{S}_t=\frac{2mC^{m-1}(1-\gamma)}{(m-\beta)^2}(-x)^{\frac{2(\beta-1)}{m-\beta}}\left[1-C^{\beta-1}(-x)^{\frac{2(\beta-1)}{m-\beta}}b(1-\beta)(1-\gamma)t\right]^{\frac{m+2\beta-3}{1-\beta}}\\
&\times\left[(m+\beta-2)(m-1)C^{\beta-1}b(1-\beta)(-x)^{\frac{2(\beta-1)}{m-\beta}}(1-\gamma)t+(m+\beta-2)(m+2\beta-1)\right]\\
&\ge\ 0 \quad \text{in } M.
\end{aligned}$$

Thus

$$\gamma - (C/C_*)^{m-\beta} \le \mathrm{S} \le \gamma \quad \text{in} \quad M.$$

If we take $\gamma = (C/C_*)^{m-\beta}$ (respectively, $\gamma = 0$), then we have

$$\mathrm{L}g \ge 0 \quad (\text{respectively, } \mathrm{L}g \le 0) \quad \text{in} \quad M, \tag{4.10a}$$

$$\mathrm{L}g = 0 \quad \text{for} \quad x > \mu_\gamma(t), \quad t > 0. \tag{4.10b}$$

As before, from Lemma 2.1 of [3] and a comparison principle, the estimation (2.13) follows.
Let $1 \le m < 2 - \beta$ and $0 < C < C_*$. First, we can establish the following rough estimation:

$$\left[C^{1-\beta}(-x)_+^{\frac{2(1-\beta)}{m-\beta}} - b(1-\beta)\Big(1 - (C/C_*)^{m-\beta}\Big)t\right]_+^{\frac{1}{1-\beta}} \le u(x,t)$$

$$\le C(-x)_+^{\frac{2}{m-\beta}} \quad \text{for} \quad x \in \mathbb{R}, \quad 0 \le t < +\infty. \tag{4.11}$$

To prove the left-hand side we consider the function g as in the case when $m + \beta > 2$ with $\gamma = (C/C_*)^{m-\beta}$. As before, we then derive (4.9a) and, since

$$\mathrm{S}_t \le 0 \quad \text{in} \quad M,$$

we have $S \le 0$ in M. Hence, (4.10) is valid with $\le$ in (4.10a). As before, from Lemma 2.1 of [3] and a comparison principle, the left-hand side of (4.11) follows. To prove the right-hand side of (4.11) it is enough to observe that

$$\mathrm{L}u_0 = bu_0^\beta\Big(1 - (C/C_*)^{m-\beta}\Big) \ge 0 \quad \text{for} \quad x \in \mathbb{R}, \quad t \ge 0,$$

and to apply the comparison principle.
Using (4.11), we can now establish a more accurate estimation (2.14). For that, consider a function

$$g(x,t) = C_0\Big(-\zeta_0\, t^{\frac{m-\beta}{2(1-\beta)}} - x\Big)_+^{\frac{2}{m-\beta}} \quad \text{in} \quad G_\ell,$$

$$G_\ell = \Big\{(x,t) : \zeta(t) = -\ell\, t^{\frac{m-\beta}{2(1-\beta)}} < x < +\infty, \quad 0 < t < +\infty\Big\},$$

where $C_0 > 0$, $\zeta_0 > 0$, $\ell > \zeta_0$ are some constants. Calculating $\mathrm{L}g$ in

$$G_\ell^+ = \Big\{(x,t) : \zeta(t) < x < -\zeta_0\, t^{\frac{m-\beta}{2(1-\beta)}}, \quad 0 < t < +\infty\Big\},$$

we have

$$\mathrm{L}g = bg^\beta \mathrm{S}, \qquad \mathrm{S} = 1 - (C_0/C_*)^{m-\beta} - \Big(b(1-\beta)\Big)^{-1} C_0^{1-\beta}\zeta_0\, t^{\frac{m+\beta-2}{2(1-\beta)}}$$

$$\times \Big(-\zeta_0\, t^{\frac{m-\beta}{2(1-\beta)}} - x\Big)^{\frac{2-m-\beta}{m-\beta}}. \tag{4.12}$$

Hence, if we take $C_0 = C_*$, then

$$\mathrm{L}g \le 0 \quad \text{in} \quad G_\ell^+; \ \mathrm{L}g = 0 \quad \text{in} \quad G_\ell \backslash \bar{G}_\ell^+. \tag{4.13}$$

To obtain a lower estimation we now choose $\zeta_0 = \zeta_3,\ \ell = \ell_0$ (see Appendix). Using (4.11), we then have

$$g(\zeta(t),t) = C_*(\ell_0 - \zeta_3)^{\frac{2}{m-\beta}} t^{\frac{1}{1-\beta}} = \Big(b(1-\beta)\theta_* t\Big)^{\frac{1}{1-\beta}}$$

$$= \left[C^{1-\beta}\ell_0^{\frac{2(1-\beta)}{m-\beta}} - b(1-\beta)\Big(1-(C/C_*)^{m-\beta}\Big)\right]^{\frac{1}{1-\beta}}\ t^{\frac{1}{1-\beta}}$$

$$\le u(\zeta(t),t), \quad t \ge 0, \tag{4.14a}$$

$$g(x,0) = u(x,0) = 0, \qquad 0 \le x \le x_0, \tag{4.14b}$$

$$g(x_0,t) = u(x_0,t) = 0, \qquad t \ge 0, \tag{4.14c}$$

where $x_0 > 0$ is an arbitrary fixed number. By using (4.13), (4.14), and Lemma 2.1 of [3], we can apply Comparison Theorem 3.4 of [3] in

$$G'_{\ell_0} = G_{\ell_0} \cap \Big\{x < x_0\Big\}.$$

Since $x_0 > 0$ is an arbitrary number the desired lower estimation from (2.14) follows.

Let us now prove the right-hand side of (2.14). Since

$$\mathrm{S}_x \ge 0 \quad \text{for} \quad \zeta(t) < x < -\zeta_0\, t^{\frac{m-\beta}{2(1-\beta)}}, \quad t > 0,$$

from (4.12) it follows that

$$\mathrm{S} \ge \mathrm{S}\Big|_{x=\zeta(t)} = 1 - (C_0/C_*)^{m-\beta} - \Big(b(1-\beta)\Big)^{-1} C_0^{1-\beta}\zeta_0(\ell-\zeta_0)^{\frac{2-m-\beta}{m-\beta}}.$$

Taking now $C_0 = C_3,\ \zeta_0 = \zeta_4,\ \ell = \ell_1$ (see Appendix), we have

$$\mathrm{S}\Big|_{x=\zeta(t)} = 0;$$

hence (by using (4.11))

$$\mathrm{L}g \ge 0 \quad \text{in} \quad G^+_{\ell_1},\ \mathrm{L}g = 0 \quad \text{in} \quad G_{\ell_1}\backslash\bar{G}^+_{\ell_1},$$

$$u(\zeta(t),t) \le C\ell_1^{\frac{2}{m-\beta}} t^{\frac{1}{1-\beta}} = C_3(\ell_1-\zeta_4)^{\frac{2}{m-\beta}} t^{\frac{1}{1-\beta}} = g(\zeta(t),t), \quad t \ge 0,$$

and, for arbitrary $x_0 > 0$, (4.14b) and (4.14c) are valid. As before, applying the Comparison Theorem 3.4 of [3] in G'_{ℓ_1}, we then derive the right-hand side of (2.14), since $x_0 > 0$ is arbitrary.

From (2.11), (2.13), and (2.14) it follows that

$$\zeta_1 t^{\frac{m-\beta}{2(1-\beta)}} \le \eta(t) \le \zeta_2 t^{\frac{m-\beta}{2(1-\beta)}}, \quad 0 \le t < +\infty,$$

where the constants ζ_1 and ζ_2 are chosen according to relevant estimations for u. However, it may easily be shown that the proved estimations (2.11), (2.13), and (2.14), together with existence, uniqueness, and comparison results for the original CP (1.1), (1.4) (see section 3), imply the unique solvability of the relevant boundary value problem for the function f_1 from (2.9), as well as the existence of a finite number ζ_* such that $\zeta_* = \sup\{\zeta : f_1(\zeta) > 0\}$. Thus (2.10) is valid. From (2.10) and the respective estimations (2.11), (2.13), or (2.14), the estimation (2.12) follows, where ζ_1 and ζ_2 are chosen according to the relevant estimation for u.

If u_0 satisfies (1.3) with $\alpha = 2/(m-\beta)$ and with $C \neq C_*$, then the asymptotic formulae (2.15) and (2.16) may be proved as the similar estimations (2.1) and (2.3) were in Lemma 3.1.

(3) Suppose that $b > 0,\ 0 < \beta < 1,\ \alpha > 2/(m-\beta),\ m \geq 1$.

Take an arbitrary sufficiently small number $\varepsilon > 0$. From (1.3), (3.5) follows. Then consider a function

$$g_\varepsilon(x,t) = \Big[(C+\varepsilon)^{1-\beta}(-x)_+^{\alpha(1-\beta)} - (1-\beta)(1-\varepsilon)t\Big]_+^{1/(1-\beta)}. \tag{4.15}$$

Let us estimate $\mathrm{L}g$ in

$$M_1 = \Big\{(x,t):\ x_\varepsilon < x < \eta_{\ell(\varepsilon)}(t), \qquad 0 < t \leq \delta_1\Big\},$$

$$\eta_\ell(t) = -\ell t^{1/\alpha(1-\beta)}, \quad \ell(\varepsilon) = (C+\varepsilon)^{-1/\alpha}\Big[b(1-\beta)(1-\varepsilon)\Big]^{1/\alpha(1-\beta)},$$

where $\delta_1 > 0$ is chosen such that $\eta_{\ell(\varepsilon)}(\delta_1) = x_\varepsilon$. We have

$$\begin{aligned}
\mathrm{L}g &= bg_\varepsilon^\beta\{\varepsilon + S\},\\
S &= -b^{-1}m\alpha(\alpha(1-\beta)-1)(C+\varepsilon)^{m-\beta}(-x)^{\alpha(m-\beta-2)}\Big\{g|x|^{-\alpha}\Big/(C+\varepsilon)\Big\}^{m-1}\\
&\quad -b^{-1}m\alpha^2(m+\beta-1)(C+\varepsilon)^{m-\beta}(-x)^{\alpha(m-\beta)-2}\Big\{g|x|^{-\alpha}\Big/(C+\varepsilon)\Big\}^{m+\beta-2}\\
&= -b^{-1}m\alpha(C+\varepsilon)^{m-\beta}(-x)^{\alpha(m-\beta)-2}\Big\{g|x|^{-\alpha}\Big/(C+\varepsilon)\Big\}^{m+\beta-2}S_1,\\
S_1 &= \Big\{(\alpha(1-\beta)-1)\Big[g|x|^{-\alpha}\Big/(C+\varepsilon)\Big]^{1-\beta} + \alpha(m+\beta-1)\Big\}.
\end{aligned}$$

If $m+\beta \geq 2$, then we can choose $x_\varepsilon < 0$ such that (with sufficiently small $|x_\varepsilon|$)

$$|S| < \varepsilon/2 \quad \text{in} \quad M_1.$$

Thus we have

$$\begin{aligned}
\mathrm{L}g_\varepsilon &> b(\varepsilon/2)g_\varepsilon^\beta \quad \Big(\text{respectively, } \mathrm{L}g_{-\varepsilon} < -b(\varepsilon/2)g_{-\varepsilon}^\beta\Big) \text{ in } M_1,\\
\mathrm{L}g_{\pm\varepsilon} &= 0 \quad \text{for } x > \eta_{\ell(\pm\varepsilon)}(t), \quad 0 < t \leq \delta_1,\\
g_\varepsilon(x,0) &\geq u_0(x) \quad \Big(\text{respectively, } g_{-\varepsilon}(x,0) \leq u_0(x)\Big), \quad x \geq x_\varepsilon.
\end{aligned}$$

Since u and g are continuous functions, $\delta = \delta(\varepsilon) \in (0, \delta_1]$ may be chosen such that

$$g_\varepsilon(x_\varepsilon,t) \geq u(x_\varepsilon,t) \quad \Big(\text{respectively, } g_{-\varepsilon}(x_\varepsilon,t) \leq u(x_\varepsilon,t)\Big), \quad 0 \leq t \leq \delta.$$

From Lemma 2.1 of [3] and Comparison Theorem 2.4 of [3] it follows that

$$g_{-\varepsilon} \leq u \leq g_\varepsilon, \qquad x \geq x_\varepsilon, \qquad 0 \leq t \leq \delta, \tag{4.16a}$$

$$\eta_{\ell(-\varepsilon)}(t) \leq \eta(t) \leq \eta_{\ell(\varepsilon)}(t), \qquad 0 \leq t \leq \delta, \tag{4.16b}$$

which imply (2.17) and (2.18).

Let $1 \le m < 2-\beta$. In this case the left-hand side of (4.16) may be proved similarly. Moreover, we can replace $1+\varepsilon$ with 1 in $g_{-\varepsilon}$ and $\eta_{\ell(-\varepsilon)}$.
To prove a relevant upper estimation, consider a function

$$g(x,t) = C_6\left(-\zeta_5 t^{\frac{1}{\alpha(1-\beta)}} - x\right)_+^{\alpha} \quad \text{in } \ G_{\ell,\delta},$$
$$G_{\ell,\delta} = \Big\{(x,t):\ \eta_\ell(t) < x < +\infty, \quad 0 < t < \delta\Big\},$$

where $\ell \in (\ell_*, +\infty)$ and

$$\zeta_5 = (\ell_*/\ell)^{\alpha(1-\beta)}(1-\varepsilon)\ell,$$
$$C_6 = \left[1-(\ell_*/\ell)^{\alpha(1-\beta)}(1-\varepsilon)\right]^{-\alpha}\left[C^{1-\beta} - \ell^{-\alpha(1-\beta)}b(1-\beta)(1-\varepsilon)\right]^{1/(1-\beta)}.$$

From (2.18) it follows that for all $\ell > \ell_*$ and for all sufficiently small $\varepsilon > 0$ there exists a $\delta = \delta(\varepsilon,\ell) > 0$ such that

$$u\Big(\eta_\ell(t), t\Big) \le \left[C^{1-\beta}\ell^{\alpha(1-\beta)} - b(1-\beta)(1-\varepsilon)\right]^{\frac{1}{1-\beta}} t^{\frac{1}{1-\beta}}, \quad 0 \le t \le \delta. \tag{4.17}$$

Calculating Lg in

$$G^+_{\ell,\delta} = \Big\{(x,t):\ \eta_\ell(t) < x < -\zeta_5 t^{\frac{1}{\alpha(1-\beta)}}, \quad 0 < t < \delta\Big\},$$

we have

$$\mathrm{L}g = bg^\beta S, \quad S = 1 - \Big(b(1-\beta)\Big)^{-1} C_6^{1/\alpha}\zeta_5\left\{gt^{1/(\beta-1)}\right\}^{1-\beta-1/\alpha}$$
$$-b^{-1}\alpha m(\alpha m-1)C_6^{2/\alpha}g^{m-\beta-2/\alpha}.$$

Since

$$S_x \ge 0 \quad \text{in } \ G^+_{\ell,\delta},$$

we have

$$S \ge S\big|_{x=\eta_\ell(t)} = 1 - \Big(b(1-\beta)\Big)^{-1} C_6^{1-\beta}\zeta_5(\ell-\zeta_5)^{\alpha(1-\beta)-1}$$
$$-b^{-1}C_6^{m-\beta}\alpha m(\alpha m-1)\left\{(\ell-\zeta_5)t^{1/\alpha(1-\beta)}\right\}^{\alpha(m-\beta)-2}.$$

Then we have

$$S \ge \varepsilon - b^{-1}C_6^{m-\beta}\alpha m(\alpha m-1)\left\{(\ell-\zeta_5)t^{1/\alpha(1-\beta)}\right\}^{\alpha(m-\beta)-2} \quad \text{in } \ G^+_{\ell,\delta}.$$

Hence, we can choose $\delta = \delta(\varepsilon) > 0$ so small that

$$\mathrm{L}g \ge b(\varepsilon/2)g^\beta \quad \text{in } \ G^+_{\ell,\delta}. \tag{4.18a}$$

Using (4.17), we can apply the Comparison Theorem 3.4 of [3] in $G'_{\ell,\delta} = G_{\ell,\delta} \cap \{x < x_0\}$, where $x_0 > 0$ is an arbitrary fixed number. We have

$$\mathrm{L}g = 0 \quad \text{in } \ G'_{\ell,\delta}\backslash\bar{G}^+_{\ell,\delta}, \tag{4.18b}$$

$$u(\eta_\ell(t),t) \le \left[C^{1-\beta}\ell^{\alpha(1-\beta)} - b(1-\beta)(1-\varepsilon)\right]^{\frac{1}{1-\beta}} t^{\frac{1}{1-\beta}}$$

$$= C_6(\ell-\zeta_5)^\alpha t^{\frac{1}{\alpha(1-\beta)}} = g(\eta_\ell(t),t), \quad 0 \le t \le \delta. \tag{4.18c}$$

$$u(x_0,t) = g(x_0,t) = 0, \qquad 0 \le t \le \delta, \tag{4.18d}$$

$$u(x,0) \ = g(x,0) \ = 0, \qquad 0 \le x \le x_0. \tag{4.18e}$$

Since $x_0 > 0$ is arbitrary, from (4.18) and the Comparison Theorem 3.4 of [3], it follows that for all $\ell > \ell_*$ and $\varepsilon > 0$ there exists $\delta = \delta(\varepsilon,\ell) > 0$ such that

$$u(x,t) \le C_6\left(-\zeta_5 t^{\frac{1}{\alpha(1-\beta)}} - x\right)_+^\alpha \quad \text{in} \quad \bar{G}_{\ell,\delta}. \tag{4.19}$$

Obviously, in view of (2.18) (which is valid along $x = \eta_\ell(t)$), δ may be chosen so small that

$$-\ell t^{1/\alpha(1-\beta)} \le \eta(t) \le -\zeta_5 t^{1/\alpha(1-\beta)}, \quad 0 \le t \le \delta. \tag{4.20}$$

Since $\ell > \ell_*$ and $\varepsilon > 0$ are arbitrary numbers, (2.17) follows from (4.20).

(4a) This case is immediate.

(4b) Let $\beta = 1$, $\alpha > 2/(m-1)$. As before, from (1.3), (3.5) follows. Then consider a function

$$g(x,t) = (C-\varepsilon)(-x)_+^\alpha \exp(-bt),$$

which satisfies

$$\mathrm{L}g \le 0 \quad \text{for} \quad x_\varepsilon < x < 0, \quad t > 0; \qquad \mathrm{L}g = 0 \quad \text{for} \quad x > 0, \quad t > 0.$$

Obviously, we can choose $\delta = \delta(\varepsilon) > 0$ such that

$$g(x_\varepsilon,t) \le u(x_\varepsilon,t), \qquad 0 \le t \le \delta_\varepsilon,$$

and from a comparison principle, the left-hand side of (2.20) immediately follows. To prove the right-hand side, consider a function

$$g(x,t) = (C+\varepsilon)(-x)_+^\alpha \exp(-bt)[1-\varepsilon(b(m-1))^{-1}(1-\exp(-b(m-1)t))]^{1/(1-m)}.$$

We have

$$\mathrm{L}g = (m-1)^{-1}(C+\varepsilon)^{1-m}(-x)^{\alpha(1-m)}\exp(-bmt)g^m$$
$$\times\left\{\varepsilon - (m-1)(C+\varepsilon)^{m-1}\alpha m(\alpha m-1)(-x)^{\alpha(m-1)-2}\right\}, \qquad x < 0, \quad t > 0,$$

and hence, if $|x_\varepsilon|$ is small enough,

$$\mathrm{L}g \ge 0 \quad \text{for} \quad x_\varepsilon < x < 0, \quad t > 0; \qquad \mathrm{L}g = 0 \quad \text{for} \quad x > 0, t > 0.$$

As before, a comparison principle implies the right-hand side of (2.20). The estimations (2.21)–(2.23) in the cases (4c) and (4d) may be proved similarly.

(II) $b = 0$.

(1) Let $m > 1$, $0 < \alpha < 2/(m-1)$.

First assume that u_0 is defined by (1.4). The self-similar form (2.25) and the formula (2.26) are well-known results (see Remark 2.1 and Lemma 3.1).

To prove (2.27), consider a function

$$g(x,t) = t^{\alpha/(2-\alpha(m-1))} f(\xi).$$

We have

$$\begin{aligned} \mathrm{L}g &= t^{(m\alpha-2)/(2-\alpha(m-1))}\, \mathcal{L}_t f, \\ \mathcal{L}_t f &= \frac{\alpha}{2-\alpha(m-1)} f - \frac{1}{2-\alpha(m-1)} \xi f' - (f^m)''. \end{aligned}$$

As a function f we take

$$f(\xi) = C_0(\xi_0 - \xi)_+^{1/(m-1)}, \qquad 0 < \xi < +\infty,$$

where C_0 and ξ_0 are some positive constants. Then we have

$$\begin{aligned} &\mathcal{L}_t f = \Big(2-\alpha(m-1)\Big)^{-1} (m-1)^{-1} C_0 (\xi_0 - \xi)^{\frac{2-m}{m-1}} R(\xi) \quad \text{for} \quad 0 \le \xi < \xi_0, \quad t > 0, \\ &R(\xi) = \alpha(m-1)\xi_0 + \Big(1-\alpha(m-1)\Big)\xi - (m-1)^{-1} m \Big(2-\alpha(m-1)\Big) C_0^{m-1}. \end{aligned}$$

To prove an upper estimation we take $C_0 = C_5,\ \xi_0 = \xi_4$. Then we have

$$R(\xi) \ge \nu_\alpha \xi_4 - (m-1)^{-1} m \Big(2-\alpha(m-1)\Big) C_5^{m-1} = 0 \text{ for } 0 \le \xi \le \xi_4,$$

where

$$\nu_\alpha = \left\{1 \ \text{ if } \ \alpha \ge (m-1)^{-1};\ \alpha(m-1) \ \text{ if } \ \alpha < (m-1)^{-1}\right\}.$$

Hence

$$\begin{aligned} \mathrm{L}g &\ge 0 \quad \text{for} \quad 0 < x < \xi_4 t^{1/(2-\alpha(m-1))}, \quad t > 0, \\ \mathrm{L}g &= 0 \quad \text{for} \quad x > \xi_4 t^{1/(2-\alpha(m-1))}, \qquad t > 0, \\ u(0,t) &= g(0,t),\ t \ge 0;\ u(x,0) = g(x,0), \qquad x \ge 0. \end{aligned}$$

Lemma 2.1 of [3] and a comparison principle imply the right-hand side of (2.27). The left-hand side of (2.27) may be established similarly if we take $C_0 = C_4,\ \xi_0 = \xi_3$. (2.2) and (2.4) are well-known results (see Remark 2.1 and Lemma 3.1). Finally, (2.5)–(2.7) easily follow from (2.26) and (2.27). If u_0 satisfies (1.3) with $0 < \alpha < 2/(m-1)$, then (2.1)–(2.3) follow from Lemma 3.1.

The cases (2) and (3) are immediate.

Appendix. We give here explicit values of the constants used in section 2 in the outline of the results for Case (I(2)) and later in section 4 during the proof of these results.

$$\zeta_1 = A_1^{\frac{m-1}{2}} \Big(m(1-\beta)\Big)^{\frac{1}{2}} \Big(1 + b(1-\beta)A_1^{\beta-1}\Big)^{-\frac{1}{2}} (m-1)^{-1}, \quad C_1 = A_1\, \zeta_1^{-\mu}$$

$$\text{if} \ \ m + \beta > 2,$$

$$\zeta_1 = A_1^{\frac{m-1}{2}}\Big(m(1-\beta)\Big)^{\frac12}\Big(1+b(1-\beta)A_1^{\beta-1}\Big)^{-\frac12}\Big(2(m+\beta)\Big)^{\frac12}(m-\beta)^{-1},$$

$$C_1 = A_1\,\zeta_1^{-\frac{2}{m-\beta}} \qquad \text{if} \quad 1 \le m < 2-\beta,$$

$$\zeta_2 = A_1^{\frac{m-1}{2}}\Big(m(1-\beta)\Big)^{\frac12}\Big(1+b(1-\beta)A_1^{\beta-1}\Big)^{-\frac12}\Big(2(m+\beta)\Big)^{\frac12}(m-\beta)^{-1},$$

$$C_2 = A_1\,\zeta_2^{-\frac{2}{m-\beta}} \qquad \text{if} \quad m+\beta > 2,$$

$$\zeta_2 = \Big(A_1/C_*\Big)^{\frac{m-\beta}{2}}, \quad C_2 = C_* \qquad \text{if} \quad 1 \le m < 2-\beta,$$

$$\bar\zeta_2 = A_1^{\frac{m-1}{2}}\left(\frac{2m(1-\beta)}{(m-\beta)(m-1)}\right)^{\frac12}, \qquad \bar C_2 = A_1\,\bar\zeta_2^{-\frac{1}{m-1}} \quad \text{if} \quad m+\beta > 2,$$

$$\bar\zeta_2 = A_1^{\frac{m-1}{2}}\Big(1+b(1-\beta)A_1^{\beta-1}\Big)^{-\frac12}\Big[m(1-\beta)\Big]^{\frac12}(m-1)^{-1}, \quad \bar C_2 = A_1\,\bar\zeta_2^{-\frac{1}{m-1}}$$
$$\text{if} \quad 1 < m < 2-\beta.$$

$$\ell_0 = C_*^{\frac{\beta-m}{2}}\Big(C_*/C\Big)^{\frac{(1-\beta)(m-\beta)}{2-m-\beta}}\Big(b(1-\beta)\theta_*\Big)^{\frac{m-\beta}{2(1-\beta)}},$$

$$\zeta_3 = C_*^{\frac{\beta-m}{2}}\left[\Big(C_*/C\Big)^{\frac{(1-\beta)(m-\beta)}{2-m-\beta}} - 1\right]\Big(b(1-\beta)\theta_*\Big)^{\frac{m-\beta}{2(1-\beta)}},$$

$$\theta_* = \left[1-\Big(C/C_*\Big)^{m-\beta}\right]\left[\Big(C_*/C\Big)^{\frac{(m-\beta)(1-\beta)}{2-m-\beta}} - 1\right]^{-1},$$

$$\ell_1 = C^{\frac{\beta-m}{2}}\left[b(1-\beta)(\delta_*\Gamma)^{-1}\Big(1-\delta_*\Gamma-(1-\delta_*\Gamma)^{-1}\Big(C/C_*\Big)^{m-\beta}\Big)\right]^{\frac{m-\beta}{2(1-\beta)}},$$

$$\zeta_4 = \delta_*\Gamma\ell_1, \qquad \Gamma = 1-\Big(C/C_*\Big)^{\frac{m-\beta}{2}}, \qquad C_3 = C\Big(1-\delta_*\Gamma\Big)^{\frac{2}{\beta-m}},$$

where $\delta_* \in (0,1)$ satisfies

$$g(\delta_*) = \max_{[0;1]} g(\delta), \qquad g(\delta) = \delta^{\frac{2-m-\beta}{m-\beta}}\left[1-\delta\Gamma-(1-\delta\Gamma)^{-1}\Big(C/C_*\Big)^{m-\beta}\right].$$

Acknowledgments. This research was done while the first author was visiting Nottingham University as a Royal Society Visiting FSU Research Fellow (March 1996–March 1997). Thanks are expressed to Mrs. Anne Perkins for the painstaking typing of this difficult manuscript.

REFERENCES

[1] U. G. Abdullaev, *Existence of unbounded solutions of a nonlinear heat equation with a sink*, Comput. Math. Math. Phys., 33 (1993), pp. 205–216.

[2] U. G. Abdullaev, *Local structure of solutions of the reaction-diffusion equations*, Nonlinear Anal., 30 (1997), pp. 3153–3163.

[3] U. G. Abdulla, *Reaction-diffusion in irregular domains*, J. Differential Equations, 164 (2000), pp. 321–354.
[4] U. G. Abdullaev, *Instantaneous shrinking and exact local estimations of solutions in nonlinear diffusion absorption*, Adv. Math. Sci. Appl., 8 (1998), pp. 483–503.
[5] H. W. Alt and S. Luckhaus, *Quasilinear elliptic-parabolic differential equations*, Math. Z., 183 (1983), pp. 311–341.
[6] L. Alvarez and J. I. Diaz, *On the initial growth of interfaces in reaction-diffusion equations with strong absorption*, Proc. Roy. Soc. Edinburgh Sect. A, 123 (1993), pp. 803–817.
[7] D. G. Aronson, L. A. Cafarelli, and S. Kamin, *How an initially stationary interface begins to move in porous medium flow*, SIAM J. Math. Anal., 14 (1983), pp. 639–658.
[8] G. I. Barenblatt, *On some unsteady motions of a liquid or a gas in a porous medium*, Prikl. Mat. Mech., 16 (1952), pp. 67–78.
[9] P. Benilan, M. G. Crandall, and M. Pierre, *Solutions of the porous medium equation in R^N under optimal conditions on initial values*, Indiana Univ. Math. J., 33 (1984), pp. 51–87.
[10] H. Brezis and A. Friedman, *Estimates on the support of solutions of parabolic variational inequalities*, Illinois J. Math., 20 (1976), pp. 82–97.
[11] L. A. Cafarelli and A. Friedman, *Continuity of the density of a gas flow in a porous medium*, Trans. Amer. Math. Soc., 252 (1979), pp. 99–113.
[12] E. DiBenedetto, *Continuity of weak solutions to a general porous medium equation*, Indiana Univ. Math. J., 32 (1983), pp. 83–118.
[13] L. C. Evans and B. F. Knerr, *Instantaneous shrinking of the support of nonnegative solutions to certain nonlinear parabolic equations and variational inequalities*, Illinois J. Math., 23 (1979), pp. 153–166.
[14] B. H Gilding and L. A. Peletier, *On a class of similarity solutions of the porous media equation*, J. Math. Anal. Appl., 55 (1976), pp. 351–364.
[15] B. H Gilding and L. A. Peletier, *On a class of similarity solution of the porous media equation* II, J. Math. Anal. Appl., 57 (1977), pp. 522–538.
[16] B. H. Gilding, *Improved theory for a nonlinear degenerate parabolic equation*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 16 (1989), pp. 165–224.
[17] R. E. Grundy and L. A. Peletier, *Short time behavior of a singular solution to the heat equation with absorption*, Proc. Roy. Soc. Edinburgh Sect. A, 107 (1987), pp. 271–288.
[18] R. E. Grundy and L. A. Peletier, *The initial interface development for a reaction-diffusion equation with power law initial data*, Quart. J. Mech. Appl. Math., 43 (1990), pp. 535–559.
[19] M. A. Herrero and J. L. Vazquez, *Thermal waves in absorbing media*, J. Differential Equations, 74 (1988), pp. 218–233.
[20] A. S. Kalashnikov, *The Cauchy problem in the class of increasing functions for equations of unsteady filtration type*, Vestnik Moskow Univ. Ser. I Mat. Mekh., 6 (1963), pp. 17–27.
[21] A. S. Kalashnikov, *The propagation of disturbances in problems of nonlinear heat conduction with absorption*, USSR Comp. Math. Math. Phys., 14 (1974), pp. 891–905.
[22] A. S. Kalashnikov, *The effect of absorption on heat propagation in a medium in which the thermal conductivity depends on temperature*, USSR Comp. Math. Math. Phys., 16 (1976), pp. 141–149.
[23] A. S. Kalashnikov, *Some problems of the qualitative theory of nonlinear degenerate second order parabolic equations*, Russian Math. Surveys, 42 (1987), pp. 169–222.
[24] S. Kamin, L. A. Peletier, and J. L. Vazquez, *A Nonlinear Diffusion-Absorption Equation with Unbounded Initial Data*, Math. Inst. Univ. of Leiden, The Netherlands, 1990, preprint.
[25] R. Kersner, *On the behavior of heat fronts in media with nonlinear heat conductivity with absorption*, Vestnik Moskow Univ. Ser. I Mat. Math., 5 (1978), pp. 44–51.
[26] R. Kersner, *Degenerate parabolic equations with general nonlinearities*, Nonlinear Anal., 4 (1980), pp. 1043–1062.
[27] J. R. King, *Development of singularities in some moving boundary problems*, Euro. J. Appl. Math., 6 (1995), pp. 491–507.
[28] B. F. Knerr, *The porous medium equation in one dimension*, Trans. Amer. Math. Soc., 234 (1977), pp. 381–415.
[29] B. F. Knerr, *The behavior of the support of solutions of the equations of nonlinear heat conduction with absorption in one dimension*, Trans. Amer. Math. Soc., 249 (1979), pp. 409–424.
[30] A. A. Lacey, J. R. Ockendon, and A. B. Tayler, *"Waiting-time" solutions of a nonlinear diffusion equation*, SIAM J. Appl. Math., 42 (1982), pp. 1252–1264.
[31] A. A. Lacey, *Initial motion of the free boundary for a nonlinear diffusion equation*, IMA J. Appl. Math., 31 (1983), pp. 113–119.

[32] O. A. Oleinik, A. S. Kalashnikov, and Y.-L Chzhou, *The Cauchy problem and boundary problems for equations of the type of non-stationary filtration*, Izv. Akad. Nauk SSSR, Ser. Mat., 22 (1958), pp. 667–704 (in Russian).

[33] A. de Pablo and J. L. Vazquez, *Travelling waves and finite propagation in a reaction-diffusion equation*, J. Differential Equations, 93 (1991), pp. 19–61.

[34] A. de Pablo and J. L. Vazquez, *The balance between strong reaction and slow diffusion*, Comm. Partial Differential Equations, 15 (1990), pp. 159–183.

[35] P. Rosenau and S. Kamin, *Thermal waves in an absorbing and convecting medium*, Phys. D, 8 (1983), pp. 273–283.

[36] J. L. Vazquez, *The interfaces of one-dimensional flows in porous media*, Trans. Amer. Math. Soc., 285 (1984), pp. 717–737.

SIAM J. MATH. ANAL.
Vol. 32, No. 2, pp. 261–290

A NOTE ON THE FORMATION OF SINGULARITIES FOR QUASI-LINEAR HYPERBOLIC SYSTEMS*

MANFRIN RENATO†

Abstract. For quasi-linear strictly hyperbolic systems, we give sufficient conditions which guarantee that singularities of the solution must occur in finite time. Moreover, we improve some of the results of [*Comm. Pure Appl. Math.*, 33 (1980), pp. 241–263] and [*Osaka J. Math.*, 34 (1997), pp. 99–113] for the wave equation $u_{tt} - a(u_x)u_{xx} = 0$. Our assumptions are rather general, in particular, we require only a weak nonlinearity condition and we do not impose restrictions on the size of the initial data.

Key words. formation of singularities, quasi-linear hyperbolic systems, weak nonlinearity

AMS subject classifications. 35L65, 35L67, 35L70

PII. S0036141098341526

Introduction. We shall be concerned with the formation of singularities of solutions for quasi-linear strictly hyperbolic 2×2 systems when the assumption of *genuine nonlinearity* (in the sense of Lax) is replaced by weaker conditions. More precisely, let us recall that the system in $\big[0,\infty\big)\times\mathbb{R}_x$

$$\begin{cases} \dfrac{\partial r}{\partial t} + \lambda(r,s)\dfrac{\partial r}{\partial x} = 0, \\[2ex] \dfrac{\partial s}{\partial t} + \mu(r,s)\dfrac{\partial s}{\partial x} = 0\,, \end{cases} \tag{1}$$

with smooth coefficients $\lambda(r,s) > \mu(r,s)$ (as is well known, general strictly hyperbolic 2×2 systems can be made diagonal, introducing the so-called *Riemann invariants*), is genuinely nonlinear if

$$\frac{\partial\lambda}{\partial r} \neq 0\,, \quad \frac{\partial\mu}{\partial s} \neq 0 \quad \text{for all } r,\, s\,. \tag{2}$$

Then, we will apply our results to the classical vibrating string equation

$$w_{tt} - a(w_x)w_{xx} = 0, \qquad w(0,x) = w_0(x),\ w_t(0,x) = w_1(x) \tag{3}$$

with $a(\eta) > 0$ smooth, which is equivalent to the system

$$\begin{aligned} &u_t - v_x = 0, \\ &v_t - a(u)u_x = 0\,, \end{aligned} \tag{4}$$

where $u = w_x$ and $v = w_t$. In fact, since $a(\eta)$ is typically an even function, we have $a'(0) = 0$ and the 2×2 system (4) is *not* genuinely nonlinear. In [10] Klainerman and Majda considered the nonlinear wave equation (3) assuming that

$$a^{(1)}(0) = \cdots = a^{(p-1)}(0) = 0, \quad a^{(p)}(0) \neq 0, \tag{5}$$

*Received by the editors July 7, 1998; accepted for publication (in revised form) July 27, 1999; published electronically June 27, 2000.
http://www.siam.org/journals/sima/32-2/34152.html
†D.C.A. Istituto Universitario di Architettura, Tolentini S. Croce 191, 30135 Venezia, Italy (manfrin@iuav.it).

for some integer $p > 1$. They proved that for sufficiently small, nontrivial (essentially) periodic initial data $w_0(x)$, $w_1(x)$, the classical C^2 solution $w(t,x)$ develops singularities in finite time.

More recently, the technique of [10] was applied by Colombini and Del Santo [2] to study the formation of singularities for system (1). When (2) is replaced by the weaker condition

$$\frac{\partial \lambda}{\partial r} \geq 0 \ (\leq 0) \quad \text{for all } (r,s) \text{ in a neighborhood } W \text{ of the origin in } \mathbb{R}^2 \tag{6}$$

and $\lambda_r(r,s)$ is not identically zero on any open subset of W, they proved the formation of singularities for small, nonconstant periodic initial data.

Then, assuming $a'(\eta) \geq 0 \ (\leq 0)$ for all $\eta \in \mathbb{R}$ and that $a(\eta)$ is not constant on any open interval, they showed that the classical C^2 solutions of (3) with small, nontrivial periodic initial data develop singularities in the second derivatives in finite time.

In this paper, without restrictions on the size of the data and under weak nonlinearity conditions, we prove that nontrivial solutions of (1) and (3) develop singularity in finite time. Our nonlinearity assumptions are of the same type considered in [2], but the class of initial data is more general.

All the proofs are based on the study of the integrals giving the *wave infinitesimal compression ratios* $\frac{\partial x_1}{\partial \alpha}(t,\alpha)$, $\frac{\partial x_2}{\partial \beta}(t,\beta)$; more precisely, let us recall that (see section 2 below)

$$\frac{\partial x_1}{\partial \alpha}(t,\alpha) = e^{h_1(t)-h_1(0)} \left\{ 1 + r_0'(\alpha) \int_0^t \lambda_r\big(r_0(\alpha), s(\tau, x_1(\tau,\alpha))\big)\, e^{h_1(0)-h_1(\tau)}\, d\tau \right\},$$

$$\frac{\partial x_2}{\partial \beta}(t,\beta) = e^{h_2(t)-h_2(0)} \left\{ 1 + s_0'(\beta) \int_0^t \mu_s\big(r(\tau, x_2(\tau,\beta)), s_0(\beta)\big)\, e^{h_2(0)-h_2(\tau)}\, d\tau \right\}, \tag{7}$$

with $\frac{\partial x_1}{\partial \alpha}(t,\alpha) > 0$, $\frac{\partial x_2}{\partial \beta}(t,\beta) > 0$ for all $t \geq 0$, $\alpha, \beta \in \mathbb{R}$, if $\big(r(t,x), s(t,x)\big)$ is a C^1 solution of (1).

In Theorem 1, we prove the formation of singularities for the C^1 solutions of (1) assuming that $\lambda_r(r,s), \mu_s(r,s) \geq 0$, with $\lambda_r(r,s) > 0$ in a dense subset of $\mathbb{R}^2$. This condition is rather restrictive, but it can be applied to the wave equation (3) generalizing the results of [10], [2] in the case $a'(\eta)$ has a zero of even order, i.e., $p = 2k+1$, $k \in \mathbb{N}$, in (5). Moreover, we give a necessary and sufficient condition for the existence of global C^2 solutions. See Theorem 2 and Corollary 5 below.

In Theorems 3 and 4 (ii) the main idea is to consider the characteristic curve $x_1(t,\alpha)$,

$$\frac{dx_1}{dt} = \lambda(r,s)(t,x_1), \quad x_1(0,\alpha) = \alpha, \tag{8}$$

for two nearby values of α; say, α_1, α_2. Then, we compare the corresponding values of the first integral in (7). More precisely, assuming that $\frac{\partial x_1}{\partial \alpha}(t,\alpha_1) > 0$ for all $t \geq 0$, we show that $\frac{\partial x_1}{\partial \alpha}(t,\alpha_2)$ goes to zero in finite time. Thus, the solution must develop a singularity along the characteristic $t \mapsto x_1(t,\alpha_2)$.

Assuming merely that $a'(\eta)\eta > 0 \ (< 0)$ for $\eta \neq 0$, in Theorem 4 (ii) we prove the formation of singularities for the *periodic* solutions of (3). This improves the results of

[10] for $p = 2k,\ k \in \mathbb{N}$, in (5). Let us remark that, in this last result, the periodicity condition plays a central role.

Finally, let us recall that the formation of singularities for 2×2 quasi-linear systems was already investigated in [1], [4], [9], [11], [15]. For $N \times N$ systems the formation of singularities (for small data) was proved by John [7] and Liu [17]. See also [6], [12], and [13],

1. Main results and remarks. Consider the following initial value problem in $\big[0,\infty\big) \times \mathbb{R}_x$:

$$\begin{cases} \dfrac{\partial r}{\partial t} + \lambda(r,s)\dfrac{\partial r}{\partial x} = 0\,, \\ \dfrac{\partial s}{\partial t} + \mu(r,s)\dfrac{\partial s}{\partial x} = 0\,, \end{cases} \tag{1.1}$$

$$r(0,x) = r_0(x), \quad s(0,x) = s_0(x), \tag{1.2}$$

where $\lambda(r,s)\,,\ \mu(r,s)$ are C^1 functions such that

$$\lambda(r,s) > \mu(r,s) \qquad \text{for all } r\,, s, \tag{1.3}$$

i.e., system (1.1) is *strictly hyperbolic*. To begin with, we will prove the following.

THEOREM 1. *Suppose that*

$$\lambda_r(r,s) \geq 0\,, \quad \mu_s(r,s) \geq 0, \tag{1.4}$$

and

$$\lambda_r(r,s) > 0 \quad \textit{in a dense subset of} \quad \mathbb{R}^2\,. \tag{1.5}$$

Moreover, assume that the initial data $r_0(x)$, $s_0(x)$ are C^1 functions satisfying the following conditions:

$$\begin{cases} \big\|r_0(x)\big\|_{C^0} < \infty\,, \qquad \big\|s_0(x)\big\|_{C^1} < \infty\,, \\ \exists\,\alpha \in \mathbb{R} \quad \textit{such that} \quad r_0'(\alpha) < 0\,, \\ \textit{there does not exist the limit} \quad \lim\limits_{x\to+\infty} s_0(x)\,. \end{cases} \tag{1.6}$$

Then, the solution of the Cauchy problem (1.1), (1.2) *must develop singularities in the first derivatives in a finite time.*

As a particular case of the above result, we have the following.

THEOREM 2. *Consider the Cauchy problem in $\big[0,\infty\big) \times \mathbb{R}_x$,*

$$\begin{aligned} &u_{tt} - a(u_x)u_{xx} = 0, \\ &u(0,x) = u_0(x), \quad u_t(0,x) = u_1(x), \end{aligned} \tag{1.7}$$

with bounded initial data $u_0(x) \in C^2\,, u_1(x) \in C^1$. Assume that $a(\eta)$ is a C^1 function such that

$$a(\eta) > 0\,, \quad a'(\eta) \geq 0 \quad \big(\textit{or } a'(\eta) \leq 0\big) \quad \text{for all } \eta \tag{1.8}$$

and $a'(\eta)$ *is not identically zero on any open interval. Then, provided* $\|u_0'(x)\|_{C^1} < \infty$, $\|u_1(x)\|_{C^1} < \infty$, *and the functions*

$$\begin{aligned} l_0(x) &= \frac{1}{2}\left\{ u_1(x) + \int_0^{u_0'(x)} a(\eta)^{1/2} d\eta \right\}, \\ r_0(x) &= \frac{1}{2}\left\{ u_1(x) - \int_0^{u_0'(x)} a(\eta)^{1/2} d\eta \right\} \end{aligned} \tag{1.9}$$

are not both monotone decreasing (*increasing*) *on* $\mathbb{R}$, *the classical* C^2 *solution* $u(t,x)$ *of* (1.7) *will develop a singularity in a finite time.*

The conditions (1.4), (1.6) can be relaxed if we assume that the maps $r \mapsto \lambda_r(r,s)$ and $s \mapsto \lambda_r(r,s)$ have discrete zeroes. More precisely, we have the following.

THEOREM 3. *Assume that* (1.3) *holds and that*

$$\begin{aligned} &\lambda_r(r,s) \geq 0, \\ &\text{the maps} \quad r \mapsto \lambda_r(r,s), \quad s \mapsto \lambda_r(r,s) \end{aligned} \tag{1.10}$$

have discrete zeroes for all s, r, *respectively. Then, provided* $r_0(x)$, $s_0(x) \in C^1$ *and*

$$\begin{cases} \|r_0(x)\|_{C^0} < \infty, \quad \|s_0(x)\|_{C^1} < \infty, \\ \exists\, \alpha_0 \in \mathbb{R} \quad \text{such that} \quad r_0'(\alpha_0) < 0, \end{cases} \tag{1.11}$$

the solution of the Cauchy problem (1.1), (1.2) *must develop singularities in finite time.*

Finally, in the last part of the paper we will prove the following.

THEOREM 4. *Consider the Cauchy problem* (1.7) *with* $u_0(x) \in C^2, u_1(x) \in C^1$ *such that*

$$\big(u_0'(x),\, u_1(x)\big) \quad \text{is a nonconstant vector.}$$

Assume that $a(\eta) > 0$ for all η *and that one of the following conditions holds*:

(i) $a(\eta)$ *is not identically constant on any open interval and*

$$u_0'(x),\, u_1(x) \to 0 \quad \text{as} \quad |x| \to \infty;$$

(ii) $u_0'(x)$, $u_1(x)$ *are periodic functions* (*with common period* $\pi > 0$) *and* $a'(\eta)$ *satisfies*

$$a'(\eta) < 0 \quad \text{for} \quad \eta < 0, \qquad a'(\eta) > 0 \quad \text{for} \quad \eta > 0. \tag{1.12}$$

Then, the solution $u(t,x)$ *of* (1.7) *will develop singularities in finite time.*

Some remarks. (1) Let us remark that, having $\lambda(r,s)$, $\mu(r,s) \in C^1$ and (1.3), the Cauchy problem (1.1), (1.2) has a *unique* local solution $\big(r(t,x),\, s(t,x)\big)$ on the stripe $\big[0,\, T\big) \times \mathbb{R}_x$ for some $T > 0$, provided

$$\|r_0(x)\|_{C^1} < \infty, \quad \|s_0(x)\|_{C^1} < \infty.$$

More generally, by the results of [5, Theorem VI], (see also [3]) we have the following.

THEOREM A (local solvability). *Consider the N-dimensional quasi-linear system in* $\big[0,\infty\big)\times\mathbb{R}_x$

$$\frac{\partial U}{\partial t} \;=\; A\big(t,x,U\big)\,\frac{\partial U}{\partial x} + B\big(t,x,U\big), \tag{1.13}$$

where $A\big(t,x,U\big)$ *is an* $N\times N$ *matrix;* U, $B\big(t,x,U\big)$ *are* N*-dimensional vectors. Besides, assume that for* $|U|$ *bounded* $A\big(t,x,U\big)$ *is a bounded* C^1 *matrix with bounded partial derivatives, while* $B\big(t,x,U\big)$ *is a bounded* C^0 *vector with continuous and bounded partial derivatives with respect to* x, U.

Finally, let us suppose that the system of (1.13) *is regularly hyperbolic. Namely, the matrix* $A\big(t,x,U\big)$ *has* N *real and distinct eigenvalues* $\lambda_1(t,x,U),\dots,\lambda_N(t,x,U)$ *such that, for* $|U|$ *bounded,*

$$\inf\ \big|\lambda_i(t,x,U)-\lambda_j(t,x,U)\big| \;>\; 0 \quad \text{for} \quad i\neq j \tag{1.14}$$

on $\big[0,\infty\big)\times\mathbb{R}_x$. *Then, the Cauchy problem*

$$\begin{aligned} &\frac{\partial U}{\partial t} \;=\; A\big(t,x,U\big)\,\frac{\partial U}{\partial x} + B\big(t,x,U\big)\,,\\ &U(0,x)=U_0(x) \qquad \text{with} \qquad \big\|U_0(x)\big\|_{C^1}<\infty \end{aligned} \tag{1.15}$$

has a unique C^1 *local solution* $U(t,x)\in C^1\big([0,T)\times\mathbb{R}_x\big)^N$, *with* $T>0$ *depending only on the* C^1 *norm of the data, i.e.,* $T=T\big(\|U_0(x)\|_{C^1}\big)$.

(2) Note that in Theorem A the life-span T depends only on the C^1 norm $\|U_0(x)\|_{C^1}$ of the initial data. Thus, if we know that $\|U(t,\cdot)\|_{C^1}$ is uniformly bounded for t bounded, then the Cauchy problem (1.15) has a unique *global* C^1 solution $U(t,x)$. See [8], [14], [18].

Now, let us consider the Cauchy problem (1.7) assuming that (1.8) holds, with $a'(\eta)$ not identically zero on any open interval, and that $\|u_0'(x)\|_{C^1}$, $\|u_1(x)\|_{C^1}<\infty$. Applying the previous arguments, we have the following conclusion.

COROLLARY 5. *The Cauchy problem* (1.7) *has a unique global classical* C^2 *solution if and only if the functions* $l_0(x)$, $r_0(x)$ *are both monotone decreasing* (*increasing*).

Proof. Thanks to Theorem 2, it is sufficient to consider the following case: $a'(\eta)\geq 0$ and $l_0(x)$, $r_0(x)$ are both monotone decreasing functions.

Introducing the Riemann invariants $l(t,x)$, $r(t,x)$ (see section 4 below), it follows that the Cauchy problem (1.7) is equivalent to the following:

$$\begin{cases} \dfrac{\partial l}{\partial t} - k(l-r)\dfrac{\partial l}{\partial x}=0, & l(0,x)=l_0(x),\\[2ex] \dfrac{\partial r}{\partial t} + k(l-r)\dfrac{\partial r}{\partial x}=0, & r(0,x)=r_0(x)\,, \end{cases} \tag{1.16}$$

where $l_0(x)$, $r_0(x)$ are the functions defined in (1.9); $k(\eta)$ is a C^1 function such that $k(\eta)>0$, $k'(\eta)\geq 0$ for all η and $k'(\eta)$ is not identically zero on any open interval.

Let $U(t,x)=\big(l(t,x),r(t,x)\big)$ be the unique (local) solution given by Theorem A. As it is well known, $l(t,x)$, $r(t,x)$ are constant along the characteristic curves (see section 2). Hence, we have

$$|l(t,x)|\leq\|l_0(x)\|_{C^0}\,,\quad |r(t,x)|\leq\|r_0(x)\|_{C^0}\,. \tag{1.17}$$

Moreover, if the initial data $l_0(x)$, $r_0(x)$ are both monotone decreasing, it is easy to prove that

$$\big|\partial_x U(t,x)\big| \le C\,\big\|\partial_x U_0(x)\big\|_{C^0} \quad \text{for all } t \ge 0 \tag{1.18}$$

for a suitable constant $C \ge 0$ independent of t. In fact, let us consider the component $l(t,x)$. By the relations (4.6), (4.7) below, we have

$$l\big(t, x_1(t,\alpha)\big) = l_0(\alpha) \quad \text{with} \quad \frac{\partial x_1}{\partial \alpha}(t,\alpha) \ge \delta > 0 \quad \text{for all } t \ge 0\,, \tag{1.19}$$

because $l_0'(x) \le 0\,, k'(\eta) \ge 0$. Hence, from (1.19) we find

$$|l_x(t,x)| \le \delta^{-1}\,\|l_0'(x)\|_{C^0}\,. \tag{1.20}$$

Clearly, in the same way it follows that $|r_x(t,x)| \le \delta^{-1}\,\|r_0'(x)\|_{C^0}$. Thus, we obtain the estimate (1.18). Finally, from (1.17), (1.18) we deduce that $\big\|U(t,\cdot)\big\|_{C^1}$ is *uniformly* bounded. Thus, applying Theorem A and the above arguments, the classical C^1 (local) solution of (1.16) exists on $\big[0,\infty\big) \times \mathbb{R}_x$. □

(3) The statements of Theorems 1, 3, 4 are still true if we "change" the inequalities. For example, case (ii) of Theorem 4 holds true if we assume that $a'(\eta) > 0$ for $\eta < 0$ and $a'(\eta) < 0$ for $\eta > 0$.

(4) In case (i) of Theorem 4 we may even suppose that $a(\eta)$ is not identically constant in any open subset of $\big(-\varepsilon, \varepsilon\big)$ for some $\varepsilon > 0$.

Notation. In the following $\|\cdot\|_{C^k}$ will denote the C^k norm, i.e.,

$$\big\|f(x)\big\|_{C^k} \stackrel{\text{def}}{=} \sum_{i=0}^{k} \sup_{x\in\mathbb{R}} \big|f^{(i)}(x)\big|\,. \tag{1.21}$$

Furthermore, the letter C will denote a generic constant, occasionally numbered for clarity.

2. The basic identities for the waves infinitesimal compression ratio. In this section, following essentially Majda [16], we compute the *waves infinitesimal compression ratio* for a smooth solution of a 2×2 strictly hyperbolic system in *one* space dimension. We give the proof in detail for convenience of the reader.

Consider the quasi-linear system in $\big[0,\infty\big) \times \mathbb{R}_x$:

$$\begin{aligned} &\frac{\partial r}{\partial t} + \lambda(r,s)\frac{\partial r}{\partial x} = 0\,,\\ &\frac{\partial s}{\partial t} + \mu(r,s)\frac{\partial s}{\partial x} = 0\,, \end{aligned} \tag{2.1}$$

where $\lambda(r,s)$, $\mu(r,s)$ are C^1 functions such that

$$\lambda(r,s) > \mu(r,s) \quad \text{for all } r,\, s\,. \tag{2.2}$$

Let $r = r(t,x)$, $s = s(t,x)$ be a smooth (C^1) bounded solution of (2.1) in $\big[0,T\big)\times\mathbb{R}_x$, with initial data

$$r(0,x) = r_0(x), \quad s(0,x) = s_0(x)\,. \tag{2.3}$$

Then, on the existence domain we define the two families of characteristic curves

$$\text{(I)} \qquad \frac{dx_1}{dt} = \lambda(r,s)(t,x_1), \quad x_1(0,\alpha) = \alpha, \tag{2.4a}$$

$$\text{(II)} \qquad \frac{dx_2}{dt} = \mu(r,s)(t,x_2), \quad x_2(0,\beta) = \beta. \tag{2.4b}$$

From (2.1), (2.3) for all $t \in \big[0,T\big)$ we have

$$r\big(t,x_1(t,\alpha)\big) = r_0(\alpha), \quad s\big(t,x_2(t,\beta)\big) = s_0(\beta), \tag{2.5}$$

i.e., $r(t,x)$ and $s(r,x)$ are constants along the graphs of $x_1(t,\alpha)$, $x_2(t,\beta)$, respectively. Differentiating both sides of (2.4a) with respect to α, we find

$$\frac{d}{dt}\Big(\frac{\partial x_1}{\partial \alpha}\Big) = \lambda_r\, r_\alpha + \lambda_s\, s_x\, \frac{\partial x_1}{\partial \alpha}\,. \tag{2.6}$$

The right-hand side of (2.6) is evaluated on the characteristic $\big(t,x_1(t,\alpha)\big)$, so we have the following identities:

$$\begin{aligned} r_\alpha &= r_0'(\alpha)\,, \\ s_x &= \frac{1}{\lambda-\mu}\,\frac{d}{dt} s\big(t,x_1(t,\alpha)\big)\,. \end{aligned} \tag{2.7}$$

Thus, defining

$$H_1(r,s) = \int_0^s \frac{\lambda_s(r,z)}{\big(\lambda-\mu\big)(r,z)} dz \tag{2.8}$$

it follows that

$$\frac{d}{dt}\Big(\frac{\partial x_1}{\partial \alpha}\Big) = \lambda_r\, r_0'(\alpha) + \frac{dH_1}{dt}\,\frac{\partial x_1}{\partial \alpha}\,. \tag{2.9}$$

Finally, integrating the linear equation (2.9) and noting that

$$\frac{\partial x_1}{\partial \alpha}(0,\alpha) = 1\,,$$

we obtain the 1-*wave infinitesimal compression ratio*:

$$\frac{\partial x_1}{\partial \alpha}(t,\alpha) = e^{h_1(t)-h_1(0)}\left\{1 + r_0'(\alpha)\int_0^t \lambda_r\big(r_0(\alpha), s(\tau,x_1(\tau,\alpha))\big)\, e^{h_1(0)-h_1(\tau)}\, d\tau\right\}, \tag{2.10}$$

where $h_1(t) = H_1\big(r(t,x_1(t,\alpha)), s(t,x_1(t,\alpha))\big) = H_1\big(r_0(\alpha), s(t,x_1(t,\alpha))\big)$.

By the same arguments, and setting

$$H_2(r,s) = \int_0^r \frac{\mu_r(z,s)}{\big(\mu-\lambda\big)(z,s)} dz\,, \tag{2.11}$$

we have the 2-*wave infinitesimal compression ratio*:

$$\frac{\partial x_2}{\partial \beta}(t,\beta) = e^{h_2(t)-h_2(0)}\left\{1 + s_0'(\beta)\int_0^t \mu_s\big(r(\tau,x_2(\tau,\beta)), s_0(\beta)\big)\, e^{h_2(0)-h_2(\tau)}\, d\tau\right\}, \tag{2.12}$$

where $h_2(t) = H_2\big(r(t,x_2(t,\beta)), s(t,x_2(t,\beta))\big) = H_2\big(r(t,x_2(t,\beta)), s_0(\beta)\big)$.

Some other remarks. In the following we will suppose $r(t,x)$, $s(t,x)$ to be a C^1 solution of (2.1) in $\big[0,T\big)\times\mathbb{R}_x$, with C^1 initial data $r_0(x)$, $s_0(x)$ such that

$$\big\|r_0(x)\big\|_{C^0}<\infty,\quad \big\|s_0(x)\big\|_{C^0}<\infty. \tag{2.13}$$

Then, we have the following.

(1) The functions $r(t,x)$, $s(t,x)$ are uniformly bounded; more precisely from (2.5) we see that

$$|r(t,x)|\le\big\|r_0(x)\big\|_{C^0},\quad |s(t,x)|\le\big\|s_0(x)\big\|_{C^0} \tag{2.14}$$

in $\big[0,T\big)\times\mathbb{R}_x$. Hence, by the assumption (2.2) and the definitions (2.8), (2.11) there exist constants $C>1$, $\delta>1$ such that

$$\begin{aligned}&\frac{1}{\delta}\le\lambda\big(r(t,x),s(t,x)\big)-\mu\big(r(t,x),s(t,x)\big)\le\delta\quad\text{on}\quad\big[0,T\big)\times\mathbb{R}_x\quad\text{and}\\ &\frac{1}{C}\le\exp\big\{h_1(t)-h_1(0)\big\}\le C\,,\\ &\frac{1}{C}\le\exp\big\{h_2(t)-h_2(0)\big\}\le C\end{aligned} \tag{2.15}$$

for all $t\in\big[0,T\big)$ and α, $\beta\in\mathbb{R}$. Here C, δ depend only on $\big\|r_0(x)\big\|_{C^0}$, $\big\|s_0(x)\big\|_{C^0}$.

(2) The *infinitesimal compression ratios* for the two waves must satisfy

$$\frac{\partial x_1}{\partial\alpha}(t,\alpha)>0,\quad \frac{\partial x_2}{\partial\beta}(t,\beta)>0 \tag{2.16}$$

for $0\le t<T$ and for all α, $\beta\in\mathbb{R}$.

In fact, assuming, for example, that

$$\frac{\partial x_1}{\partial\alpha}(t_0,\alpha_0)=0\,, \tag{2.17}$$

for some $t_0\in(0,T)$ and $\alpha_0\in\mathbb{R}$, then we can easily find a contradiction.

From (2.10), (2.17) it follows that

$$r_0'(\alpha_0)\neq 0\,. \tag{2.18}$$

On the other hand, by the relation (2.5) we know that $r\big(t,x_1(t,\alpha_0)\big)=r_0(\alpha_0)$; hence having

$$r_0'(\alpha_0)=\frac{\partial r}{\partial x}\big(t,x_1(t,\alpha_0)\big)\ \frac{\partial x_1}{\partial\alpha}(t,\alpha_0)\,, \tag{2.19}$$

it follows that $r_0'(\alpha_0)=0$, because the right-hand side of (2.19) vanishes for $t=t_0$.

(3) For any fixed $\bar\alpha\in\mathbb{R}$ ($\bar\beta\in\mathbb{R}$) the equation

$$x_1(t,\bar\alpha)=x_2(t,\beta),\quad \big(x_1(t,\alpha)=x_2(t,\bar\beta)\big) \tag{2.20}$$

has a unique solution $\beta=\beta(t)$ ($\alpha=\alpha(t)$). The function $t\mapsto\beta(t)$ is C^1, strictly increasing, and verifies the relation

$$\frac{d\beta(t)}{dt}=\frac{\big[\lambda(r,s)-\mu(r,s)\big](t,x_1(t,\bar\alpha))}{\dfrac{\partial x_2}{\partial\beta}(t,\beta(t))}\,. \tag{2.21}$$

Moreover, if $\|s_0'(x)\|_{C^0} < \infty$, then for some constant $C > 1$ we have

$$\bar\alpha + \frac{1}{C}\ln(1+t) \le \beta(t) \le \bar\alpha + C\,t \quad \text{in} \quad [0,T). \tag{2.22}$$

Proof. The existence of a unique solution $\beta = \beta(t)$ of (2.20) is a consequence of the fact that for any $t \in [0,T)$ the Cauchy problem

$$\frac{dx}{d\tau} = \mu(r,s)(\tau,x), \quad x(\tau)\Big|_{\tau=t} = x_1(t,\bar\alpha) \tag{2.23}$$

has a unique solution in $x(\tau,t)$ defined for $\tau \in [0,T)$. Obviously, we set

$$\beta(t) = x(0,t) \tag{2.24}$$

and, since $\lambda(r,s) > \mu(r,s)$, we find that $\beta(t) \ge \bar\alpha$. Then, the relation (2.21) follows from (2) applying the implicit function theorem to (2.20).

To prove the second estimate in (2.22), it is sufficient to observe that

$$x_1(t,\bar\alpha) \le \bar\alpha + t\,\max\big|\lambda(r,s)\big|, \quad \beta(t) = x(0,t) \le x_1(t,\bar\alpha) + t\,\max\big|\mu(r,s)\big| \tag{2.25}$$

for $r = r(t,x)$, $s = s(t,x)$. Finally, from (2.21) and the inequalities of (2.15), we have

$$\frac{d\beta}{dt} \ge \frac{\delta^{-1}}{\dfrac{\partial x_2}{\partial\beta}(t,\beta(t))} \ge \frac{1}{C(1+t)} \tag{2.26}$$

because, by (2.12) and the last inequality of (2.15),

$$\frac{\partial x_2}{\partial\beta}(t,\beta) \le C\Big(1 + C\,|s_0'(\beta)|\,\max|\mu_s(r,s)|\,t\Big). \tag{2.27}$$

Thus, integrating (2.26) we obtain the first estimate in (2.22). ∎

(4) For all t_a, $t_b \ge 0$ and for all $\alpha \in \mathbb{R}$, we have

$$\begin{aligned} \frac{1}{C}\,\frac{\partial x_1}{\partial\alpha}(t_b,\alpha) - C\,|r_0'(\alpha)|\,\big|t_b - t_a\big| \;&\le\; \frac{\partial x_1}{\partial\alpha}(t_a,\alpha) \\ &\le\; C\,\frac{\partial x_1}{\partial\alpha}(t_b,\alpha) + C\,|r_0'(\alpha)|\,\big|t_b - t_a\big|, \end{aligned} \tag{2.28}$$

where $C > 1$ is suitable constant, independent of t_a, t_b, α.

Clearly, a similar estimate holds true for the 2-*wave infinitesimal compression ratio.*

Proof. It is sufficient to consider the expression (2.10) of $\frac{\partial x_1}{\partial\alpha}(t,\alpha)$. In fact, we have

$$\begin{aligned} \frac{\partial x_1}{\partial\alpha}(t_a,\alpha) \;&=\; e^{h_1(t_a)-h_1(t_b)}\,\frac{\partial x_1}{\partial\alpha}(t_b,\alpha) \\ &\quad + r_0'(\alpha)\int_{t_b}^{t_a} \lambda_r\big(r_0(\alpha), s(\tau,x_1(\tau,\alpha))\big)\,e^{h_1(t_a)-h_1(\tau)}\,d\tau\,. \end{aligned} \tag{2.29}$$

Then, using the estimates of (2.15), we obtain (2.28). ∎

3. Proof of Theorem 1. We argue by contradiction. Let $r = r(t,x)$, $s = s(t,x)$ be a C^1 solution to problem (1.1), (1.2) on $\big[0,\infty\big)\times\mathbb{R}_x$. From the assumptions (1.4), (1.5), (1.6) on $\lambda(r,s)$, $\mu(r,s)$, and $s_0(x)$ we can find $\bar\alpha$, $\bar s \in \mathbb{R}$ such that

$$\begin{aligned} &r_0'(\bar\alpha) < 0\,, \\ &\lambda_r\big(r_0(\bar\alpha), s\big) \geq \epsilon > 0 \qquad \text{for} \quad \big|s-\bar s\big| < \eta\,, \end{aligned} \tag{3.1}$$

and

$$\liminf_{x\to+\infty} s_0(x) + 2\eta < \bar s < \limsup_{x\to+\infty} s_0(x) - 2\eta\,, \tag{3.2}$$

for suitable ϵ, $\eta > 0$. Hence, in formula (2.10) for $\frac{\partial x_1}{\partial\alpha}$ with $\alpha = \bar\alpha$, we will have

$$\lambda(\tau) = \lambda_r\big(r_0(\bar\alpha), s(\tau, x_1(\tau,\bar\alpha))\big) \geq \epsilon \tag{3.3}$$

when $\big|s(\tau, x_1(\tau,\bar\alpha)) - \bar s\big| \leq \eta$ and

$$\frac{\partial x_1}{\partial\alpha}(t,\bar\alpha) \leq e^{h_1(t)-h_1(0)}\left(1 + r_0'(\bar\alpha)\,\frac{\epsilon}{C}\int_0^t \chi_{\bar\alpha,\epsilon}(\tau)\,d\tau\right), \tag{3.4}$$

where $\chi_{\bar\alpha,\epsilon}(t)$ is the characteristic function of the set

$$\Omega_{\bar\alpha,\epsilon} = \Big\{t \geq 0 \,\Big|\, \lambda_r\big(r_0(\bar\alpha), s(t, x_1(t,\bar\alpha))\big) \geq \epsilon\Big\}. \tag{3.5}$$

Now, let us consider the equation

$$x_1(t,\bar\alpha) = x_2(t,\beta) \quad \text{for} \quad t \in \big[0,\infty\big)\,. \tag{3.6}$$

From the remarks of the previous section, we know that the unique solution $\beta = \beta(t)$ of (3.6) is a C^1 strictly increasing function and (since $\big\|s_0(x)\big\|_{C^1} < \infty$)

$$\beta(0) = \bar\alpha\,, \quad \beta(t) \geq \bar\alpha + \frac{1}{C}\,\ln(1+t) \tag{3.7}$$

for all $t \geq 0$. Besides, by (2.12) and having $\mu_s(r,s) \geq 0$,

$$s_0'(\beta) \geq 0 \;\Rightarrow\; \frac{\partial x_2}{\partial\beta}(t,\beta) \geq \frac{1}{C}\;; \tag{3.8}$$

hence applying (2.21) and the first inequality of (2.15), we obtain that

$$0 < \frac{d\beta(t)}{dt} = \frac{\big(\lambda(r,s) - \mu(r,s)\big)(t, x_1(t,\bar\alpha))}{\dfrac{\partial x_2}{\partial\beta}(t,\beta(t))} \leq \delta\,C\,, \tag{3.9}$$

when $s_0'(\beta(t)) \geq 0$. To conclude, it is sufficient to observe that the conditions on $s_0(x)$, that is $\big\|s_0(x)\big\|_{C^1} < \infty$ and (3.2), imply that

$$B_{\bar s,\eta} = \Big\{x \geq \bar\alpha \,\Big|\, s_0'(x) \geq 0\,, \quad \big|\bar s - s_0(x)\big| \leq \eta\Big\} \tag{3.10}$$

has *infinite* Lebesgue measure, i.e.,

$$\text{meas}\Big\{B_{\bar{s},\eta}\Big\} = +\infty\,. \tag{3.11}$$

But, for all $t \geq 0$

$$\text{meas}\Big\{B_{\bar{s},\eta} \cap \big[0,\beta(t)\big]\Big\} \leq \int_0^t \frac{d\beta(\tau)}{dt}\chi_{\bar{\alpha},\epsilon}(\tau)\,\chi_0(\tau)\,d\tau \leq \delta\,C\,\int_0^t \chi_{\bar{\alpha},\epsilon}(\tau)\,d\tau\,, \tag{3.12}$$

where $\chi_0(t)$ is the characteristic function of the set $\big\{t\,:\,s_0'\big(\beta(t)\big) \geq 0\,\big\}$.

Thus returning to (3.4), we obtain that there exists $\tilde{T}$, $0 < \tilde{T} < \infty$, such that

$$\frac{\partial x_1}{\partial \alpha}(\tilde{T},\bar{\alpha}) = 0, \tag{3.13}$$

and having

$$\frac{\partial r}{\partial x}(t,x_1(t,\bar{\alpha})) = r_0'(\bar{\alpha})\left(\frac{\partial x_1}{\partial \alpha}(t,\bar{\alpha})\right)^{-1} \tag{3.14}$$

we deduce that the function $r(t,x)$ must develop a singularity along the characteristic $\big(t,x_1(t,\bar{\alpha})\big)$ for some $t\,, 0 < t \leq \tilde{T}$. This completes the proof of Theorem 1. □

4. Proof of Theorem 2. As it is known the quasi-linear wave equation (1.7) can be reduced, introducing the Riemann invariants, to the first order diagonal system

$$\begin{aligned} &\frac{\partial l}{\partial t} - k(l-r)\frac{\partial l}{\partial x} = 0,\\ &\frac{\partial r}{\partial t} + k(l-r)\frac{\partial r}{\partial x} = 0\,, \end{aligned} \tag{4.1}$$

where

$$\begin{aligned} l &= \frac{1}{2}\left\{u_t + \int_0^{u_x} a(\eta)^{1/2}d\eta\right\},\\ r &= \frac{1}{2}\left\{u_t - \int_0^{u_x} a(\eta)^{1/2}d\eta\right\}, \end{aligned} \tag{4.2}$$

and $k(\eta) = a\big(M^{-1}(\eta)\big)^{1/2}$ with $M(\omega) = \int_0^\omega a(\eta)^{1/2}d\eta$. See [10].

Then, thanks to the assumptions (1.8), we have

$$k'(\eta) = \frac{1}{2}\,\frac{a'\big(M^{-1}(\eta)\big)}{a\big(M^{-1}(\eta)\big)} \geq 0 \qquad \big(\text{or } k'(\eta) \leq 0 \text{ if } a'(\eta) \leq 0\big) \tag{4.3}$$

and $k'(\eta)$ is not identically zero on any open interval. Moreover, by the hypotheses on the initial data, we have that $\big\|l_0(x)\big\|_{C^1}\,,\ \big\|r_0(x)\big\|_{C^1} < \infty$ and the function $l_0(x), r_0(x)$ are not both monotone decreasing.

Assume, for instance, that there exists $\alpha \in \mathbb{R}$ such that

$$l_0'(\alpha) > 0\,.$$

Then we have two possibilities:

$$\begin{aligned} &\text{(a)} \quad \text{there does not exist the limit} \quad \lim_{x\to-\infty} r_0(x)\,,\\ &\text{(b)} \quad \lim_{x\to-\infty} r_0(x) = r \in \mathbb{R}\,. \end{aligned} \tag{4.4}$$

In case (a) it is sufficient to apply Theorem 1.

In case (b), we can find $\bar\alpha$ close to α, such that

$$l_0'(\bar\alpha) > 0\,, \quad k'\big(l_0(\bar\alpha) - r\big) > 0\,. \tag{4.5}$$

Then we consider the characteristic

$$\frac{dx_1}{dt} = -k(l-r)(t,x_1), \quad x_1(0,\bar\alpha) = \bar\alpha\,. \tag{4.6}$$

Applying the 1-*wave infinitesimal compression ratio* (2.10), we have

$$\frac{\partial x_1}{\partial\alpha}(t,\bar\alpha) = e^{h_1(t)-h_1(0)}\left\{1 - l_0'(\bar\alpha)\int_0^t k'\big(l_0(\bar\alpha) - r(\tau,x_1(\tau,\bar\alpha))\big)\,e^{h_1(0)-h_1(\tau)}d\tau\right\}\,, \tag{4.7}$$

with $h_1(t) = H_1\big(l_0(\bar\alpha), r(t,x_1(t,\bar\alpha))\big)$ and

$$\lim_{t\to+\infty} r\big(t,x_1(t,\bar\alpha)\big) = r\,. \tag{4.8}$$

In fact $r\big(t,x_1(t,\bar\alpha)\big) = r_0(\beta)$, where β verifies the relation

$$x_1(t,\bar\alpha) = x_2(t,\beta)\,, \tag{4.9}$$

with $x_2(t,\beta)$ the characteristic curve defined by

$$\frac{dx_2}{dt} = k(l-r)(t,x_2), \quad x_2(0,\beta) = \beta\,. \tag{4.10}$$

But, since the data $l_0(x)$, $r_0(x)$ are bounded, there exists $\delta > 0$ such that

$$k(r-l)(t,x) \geq \delta \quad \text{for all } t,\, x\,. \tag{4.11}$$

It follows that the *unique* solution $\beta = \beta(t)$ of (4.9) satisfies

$$\beta(t) \leq \bar\alpha - 2\,\delta\,t \tag{4.12}$$

and thus (4.8) holds. Hence, from (4.7) we deduce that $\frac{\partial x_1}{\partial\alpha}(t,\bar\alpha)$ goes to zero in a finite time. This concludes the proof of Theorem 2. □

5. Proof of Theorem 3. Assume that (1.10), (1.11) hold and, by contradiction, let $r(t,x)$, $s(t,x)$ be a C^1 global solution in $\big[0,\infty\big)\times\mathbb{R}_x$ of the Cauchy problem (1.1), (1.2). Then, recalling the considerations of section 2, we must have

$$\frac{\partial x_1}{\partial\alpha}(t,\alpha) > 0\,, \quad \frac{\partial x_2}{\partial\beta}(t,\beta) > 0 \tag{2.16}$$

for all $t \geq 0$ and for all $\alpha,\,\beta \in \mathbb{R}$. We will prove that the first inequality of (2.16) does not hold.

To begin with, from (1.11) there exists $\alpha_0 \in \mathbb{R}$ such that $r_0'(\alpha_0) < 0$. Then $\frac{\partial x_1}{\partial\alpha}(t,\alpha_0) > 0$ for all $t \geq 0$ and the 1-*wave infinitesimal compression ratio* (2.10) gives the inequality

$$\int_0^t \lambda_r\big(r_0(\alpha_0), s(\tau,x_1(\tau,\alpha_0))\big)e^{h_1(0)-h_1(\tau)}\,d\tau \;<\; \frac{-1}{r_0'(\alpha_0)} \quad \text{for all } t \geq 0\,, \tag{5.1}$$

where $h_1(t) = H_1\big(r_0(\alpha_0), s(t, x_1(t, \alpha_0))\big)$.

Moreover, for the fixed α_0 and for $\beta \geq \alpha_0$, we can solve the equation

$$x_1(t, \alpha_0) = x_2(t, \beta) \tag{5.2}$$

with respect to t; using (2.21), (2.22) we have $t = t(\beta)$ with $t(\beta) : \big[\alpha_0, \infty\big) \to \infty$ a C^1 strictly increasing function such that

$$\frac{dt}{d\beta} = \frac{\dfrac{\partial x_2}{\partial \beta}(t(\beta), \beta)}{\big[\lambda(r,s) - \mu(r,s)\big](t(\beta), x_1(t(\beta), \alpha_0))}\,. \tag{5.3}$$

Hence, we can rewrite (5.1) in the following form:

(5.4)

$$\int_{\alpha_0}^{\bar\beta} \frac{\lambda_r\big(r_0(\alpha_0), s_0(\beta)\big)\,\dfrac{\partial x_2}{\partial \beta}(t(\beta), \beta)\, e^{h_1(0) - h_1(t(\beta))}}{\big[\lambda(r,s) - \mu(r,s)\big](t(\beta), x_1(t(\beta), \alpha_0))}\, d\beta \,<\, \frac{-1}{r_0'(\alpha_0)} \quad \text{for all } \bar\beta \geq \alpha_0\,.$$

Then, having $\lambda_r(r,s) \geq 0$ and $\frac{\partial x_2}{\partial \beta} > 0$, the inequalities of (2.15) give

$$\int_{\alpha_0}^{\bar\beta} \lambda_r\big(r_0(\alpha_0), s_0(\beta)\big)\,\frac{\partial x_2}{\partial \beta}(t(\beta), \beta)\, d\beta \,< C \quad \text{for all } \bar\beta \geq \alpha_0 \tag{5.5}$$

for a suitable constant $C > 0$ depending only on $\big\|r_0(x)\big\|_{C^0}$, $\big\|s_0(x)\big\|_{C^0}$ and $r_0'(\alpha_0)$.

On the other hand, by (2.22)

$$t(\beta) \geq \frac{1}{C}\big(\beta - \alpha_0\big) \quad \text{for} \quad \beta \geq \alpha_0 \tag{5.6}$$

and, using (2.15), (5.3), we deduce that

$$\int_{\alpha_0}^{\bar\beta} \frac{\partial x_2}{\partial \beta}(t(\beta), \beta)\, d\beta \,\geq\, \frac{1}{C_1}\big(\bar\beta - \alpha_0\big) \quad \text{for all } \bar\beta \geq \alpha_0\,. \tag{5.7}$$

Then, defining for $\varepsilon > 0$

$$D_\epsilon = \Big\{\beta \geq \alpha_0 \,\Big|\, \lambda_r\big(r(\alpha_0), s_0(\beta)\big) \leq \varepsilon\Big\}\,, \tag{5.8}$$

the inequalities (5.5), (5.7) give

$$\int_{\alpha_0}^{\bar\beta} \frac{\partial x_2}{\partial \beta}(t(\beta), \beta)\, \chi_{D_\varepsilon}(\beta)\, d\beta \,\geq\, \frac{1}{C_1}\big(\bar\beta - \alpha_0\big) - \frac{C}{\varepsilon} \quad \text{for all } \bar\beta \geq \alpha_0\,, \tag{5.9}$$

where $\chi_{D_\varepsilon}(\beta)$ is the characteristic function of the set D_ϵ.

Now, let us consider α_1 close to α_0 such that

$$r_0'(\alpha) < 0 \quad \text{in} \quad \big[\alpha_0, \alpha_1\big]. \tag{5.10}$$

Solving for $\beta \geq \beta_0 = \max\big\{\alpha_0,\, \alpha_1\big\}$ the equation

$$x_1(t, \alpha_1) = x_2(t, \beta) \tag{5.11}$$

with respect to t, we obtain $t = t_1(\beta)$ with

$$\big|t_1(\beta) - t(\beta)\big| \le C\,\big|\alpha_1 - \alpha_0\big|\,. \tag{5.12}$$

To prove the estimate (5.12), it is sufficient to observe that for any fixed $\beta \in \mathbb{R}$ the equation

$$x_1(t,\alpha) = x_2(t,\beta)$$

implies that

$$\frac{\partial t}{\partial \alpha} = \frac{\dfrac{\partial x_1}{\partial \alpha}(t,\alpha)}{\big[\mu(r,s) - \lambda(r,s)\big](t, x_2(t,\beta))}\,;$$

hence, having

$$0 < \frac{\partial x_1}{\partial \alpha}(t,\alpha) \le e^{h_1(t) - h_1(0)} \quad \text{for} \quad \alpha \in \big[\alpha_0,\, \alpha_1\big] \tag{5.13}$$

since $r_0'(\alpha) \le 0$ in $\big[\alpha_0,\, \alpha_1\big]$, from (2.15) we conclude that

$$\Big|\frac{\partial t}{\partial \alpha}\Big| \le C \quad \text{in} \quad \big[\alpha_0, \alpha_1\big] \tag{5.14}$$

for some $C > 0$, uniformly with respect to $\beta \ge \beta_0$.

To continue, let us prove the following.

LEMMA 1. *For $\big|\alpha_1 - \alpha_0\big|$ small enough, for all $\bar\beta \ge \beta_0 = \max\big\{\alpha_0,\, \alpha_1\big\}$ we have*

$$\int_{\beta_0}^{\bar\beta} \frac{\partial x_2}{\partial \beta}\big(t_1(\beta),\beta\big)\,\chi_{D_\varepsilon}(\beta)\,d\beta \ge \frac{\bar\beta}{C} - C\big(|\beta_0| + |\alpha_0|\big) - \frac{C}{\varepsilon} \tag{5.15}$$

with $C > 1$ independent of $\bar\beta$, ε. Moreover, let us remark that (5.15) *holds true provided $\big|\alpha_1 - \alpha_0\big|$ is sufficiently small, independently of $\varepsilon > 0$.*

Proof. To begin with, let us denote with Ω the subset of $\beta \ge \beta_0 = \max\big\{\alpha_0,\, \alpha_1\big\}$ such that

$$s_0'(\beta) \int_0^{t(\beta)} \mu_s\big(r(\tau, x_2(\tau,\beta)), s_0(\beta)\big)\, e^{h_2(0) - h_2(\tau)}\,d\tau \;\le\; 1\,. \tag{5.16}$$

Then, we can easily see that

$$\frac{\partial x_2}{\partial \beta}\big(t_1(\beta),\beta\big) \;\to\; \frac{\partial x_2}{\partial \beta}\big(t(\beta),\beta\big) \tag{5.17}$$

uniformly in Ω as $\alpha_1 \to \alpha_0$. In fact

$$\frac{\partial x_2}{\partial \beta}\big(t,\beta\big) = e^{h_2(t) - h_2(0)}\left\{1 + s_0'(\beta)\int_0^t \mu_s\big(r(\tau, x_2(\tau,\beta)), s_0(\beta)\big) e^{h_2(0) - h_2(\tau)}\,d\tau\right\}, \tag{5.18}$$

where, by the assumptions on $r_0(x)$ and $s_0(x)$, we know that the terms

$$s_0'(\beta), \quad \mu_s\big(r(\tau, x_2(\tau,\beta)), s_0(\beta)\big), \quad h_2(t)$$

are uniformly bounded. Moreover, having $h_2(t) = H_2\big(r(t, x_2(t,\beta)), s_0(\beta)\big)$, it follows that

$$h_2\big(t_1(\beta)\big) = H_2\big(r_0(\alpha_1), s_0(\beta)\big)\,. \tag{5.19}$$

Thus $h_2\big(t_1(\beta)\big) \to h_2\big(t(\beta)\big) = H_2\big(r_0(\alpha_0), s_0(\beta)\big)$ uniformly as $\alpha_1 \to \alpha_0$. Then, using (5.12) and the condition (5.16), we immediately get (5.17) in Ω.

On the other hand, for $\beta \geq \beta_0 = \max\big\{\alpha_0,\, \alpha_1\big\}$, $\beta \notin \Omega$, we must have

$$s_0'(\beta) \int_0^{t(\beta)} \mu_s\big(r(\tau, x_2(\tau,\beta)), s_0(\beta)\big)\, e^{h_2(0)-h_2(\tau)}\, d\tau > 1; \tag{5.20}$$

thus from (5.18) and the previous considerations, we deduce that

$$\frac{\partial x_2}{\partial \beta}\big(t_1(\beta),\beta\big) \geq \frac{1}{2}\,\frac{\partial x_2}{\partial \beta}\big(t(\beta),\beta\big) \quad \text{for} \quad \beta \geq \beta_0,\ \beta \notin \Omega \tag{5.21}$$

provided $\big|\alpha_1 - \alpha_0\big|$ is sufficiently small.

Taking account of these facts, we will have

$$\begin{aligned}\int_{\beta_0}^{\bar\beta} \frac{\partial x_2}{\partial \beta}\big(t_1(\beta),\beta\big)\,\chi_{D_\varepsilon}(\beta)\,d\beta &\geq \frac{1}{2}\int_{\beta_0}^{\bar\beta} \frac{\partial x_2}{\partial \beta}\big(t(\beta),\beta\big)\,\chi_{D_\varepsilon}(\beta)\,\big(1-\chi_\Omega(\beta)\big)\,d\beta \\ &+ \int_{\beta_0}^{\bar\beta} \frac{\partial x_2}{\partial \beta}\big(t_1(\beta),\beta\big)\,\chi_{D_\varepsilon}(\beta)\,\chi_\Omega(\beta)\,d\beta \\ &\geq \frac{1}{2}\int_{\alpha_0}^{\bar\beta} \frac{\partial x_2}{\partial \beta}\big(t(\beta),\beta\big)\,\chi_{D_\varepsilon}(\beta)\,d\beta - \frac{1}{4C_1}\big(\bar\beta-\beta_0\big) - C\big(\beta_0-\alpha_0\big)\end{aligned} \tag{5.22}$$

for $\big|\alpha_1-\alpha_0\big|$ sufficiently small, with C_1 the same constant appearing in (5.9). Now, from (5.9) and (5.22) we easily obtain the estimate (5.15) with a suitable constant $C > 1$. This completes the proof of Lemma 1. □

Finally, to conclude the proof of Theorem 3, we use the second assumption in (1.10). For α_1 sufficiently close to α_0 the sets

$$\begin{aligned}Z_0 &= \Big\{ s \,\Big|\, \lambda_r\big(r_0(\alpha_0), s\big) = 0,\ |s| \leq \big\|s_0(x)\big\|_{C^0} + 1 \Big\}\,, \\ Z_1 &= \Big\{ s \,\Big|\, \lambda_r\big(r_0(\alpha_1), s\big) = 0,\ |s| \leq \big\|s_0(x)\big\|_{C^0} + 1 \Big\}\end{aligned} \tag{5.23}$$

are finite and disjoint, i.e.,

$$Z_0 \cap Z_1 \,=\, \emptyset\,. \tag{5.24}$$

Hence, taking $\varepsilon > 0$ sufficiently small in the definition (5.8) of D_ε, we have

$$\lambda_r\big(r_0(\alpha_1), s_0(\beta)\big) \geq \eta \quad \text{for all } \beta \in D_\varepsilon \tag{5.25}$$

for some $\eta > 0$. Then, putting together (5.15) and (5.25), we have

$$\begin{aligned}\int_{\beta_0}^{\bar\beta} \lambda_r\big(r_0(\alpha_1), s_0(\beta)\big)\,\frac{\partial x_2}{\partial \beta}\big(t_1(\beta),\beta\big)\,d\beta &\geq \eta \int_{\beta_0}^{\bar\beta} \frac{\partial x_2}{\partial \beta}\big(t_1(\beta),\beta\big)\,\chi_{D_\varepsilon}(\beta)\,d\beta \\ &\geq \frac{\eta\bar\beta}{C} - \eta C\big(|\beta_0| + |\alpha_0|\big) - \frac{\eta\, C}{2\varepsilon}\,.\end{aligned} \tag{5.26}$$

Clearly, the last inequality and $r_0'(\alpha_1) < 0$ imply that

$$\frac{\partial x_1}{\partial \alpha}(t, \alpha_1) \to -\infty \tag{5.27}$$

as $t \to \infty$. This proves Theorem 3. □

6. Proof of Theorem 4.

Case (i). By assumption $u_0'(x)$, $u_1(x)$ are not both constant functions and

$$\lim_{x \to \pm\infty} u_0'(x) = \lim_{x \to \pm\infty} u_1(x) = 0\,. \tag{6.1}$$

We will use the Riemann invariants $l(t,x)$, $r(t,x)$ introduced in section 4. See (4.1), (4.2). From (4.3), we find that

$$k'(\eta) = \frac{1}{2}\,\frac{a'\big(M^{-1}(\eta)\big)}{a\big(M^{-1}(\eta)\big)} \tag{6.2}$$

is not identically 0 on any open interval.

Let us suppose that $l_0(x)$ is not identically 0. Then, we can find $\bar\alpha \in \mathbb{R}$ such that

$$l_0'(\bar\alpha)\, k'\big(l_0(\bar\alpha)\big) > 0\,. \tag{6.3}$$

But, rewriting (4.7), we have

$$\frac{\partial x_1}{\partial \alpha}(t, \bar\alpha) = e^{h_1(t) - h_1(0)} \left\{ 1 - l_0'(\bar\alpha) \int_0^t k'\big(l_0(\bar\alpha) - r(\tau, x_1(\tau, \bar\alpha))\big)\, e^{h_1(0) - h_1(\tau)} d\tau \right\}, \tag{6.4}$$

where

$$k'\big(l_0(\bar\alpha) - r(t, x_1(t, \bar\alpha))\big) \to k'\big(l_0(\bar\alpha)\big) \quad \text{as} \quad t \to +\infty\,. \tag{6.5}$$

In fact, the initial data $l_0(x)$, $r_0(x)$ are uniformly bounded.

Hence, by the arguments of (4.9)–(4.12), we may prove that $r\big(t, x_1(t, \bar\alpha)\big) = r_0(\beta)$, where $\beta = \beta(t)$ satisfies

$$\beta(t) \le \bar\alpha - 2\,\delta\, t \qquad (\delta > 0). \tag{6.6}$$

Thus $r\big(t, x_1(t, \bar\alpha)\big) \to 0$ as $t \to \infty$ and $\frac{\partial x_1}{\partial \alpha}(t, \bar\alpha)$ goes to 0 in finite time. We can proceed in the same way if

$$r_0(x) \not\equiv 0\,.$$

This completes the proof of Case (i) of Theorem 4. □

Case (ii). By contradiction, let $l(t,x)$, $r(t,x)$ be a global C^1 solution of the system (4.1) of the Riemann invariants.

Now, from (1.12) and (6.2), we have

$$k'(\eta) < 0 \quad \text{if} \quad \eta < 0\,, \quad k'(\eta) > 0 \quad \text{if} \quad \eta > 0\,.$$

Clearly, if

$$\max l_0(x) \ne \max r_0(x) \quad \text{or} \quad \min l_0(x) \ne \min r_0(x)\,,$$

then we can proceed as in Case (i): assuming (for example) that $\max l_0(x) > \max r_0(x)$, thanks to the condition (1.12), we can find $\bar\alpha \in \mathbb{R}$ such that

$$l_0'(\bar\alpha) > 0\,, \quad k'\big(l_0(\bar\alpha) - r(t, x_1(t, \bar\alpha))\big) \geq \epsilon > 0 \tag{6.7}$$

for all $t \geq 0$.

Hence, in the following we may assume that $l_0(x)$, $r_0(x)$ are nonconstant periodic functions and that

$$M = \max l_0(x) = \max r_0(x) > \min l_0(x) = \min r_0(x) = m\,. \tag{6.8}$$

We will denote with $\pi > 0$ the common period of the initial data $l_0(x)$, $r_0(x)$.

To begin with, rewriting the system (4.1),

$$\begin{aligned} &\frac{\partial l}{\partial t} - k(l-r)\frac{\partial l}{\partial x} = 0,\\ &\frac{\partial r}{\partial t} + k(l-r)\frac{\partial r}{\partial x} = 0\,, \end{aligned} \tag{4.1}$$

we know that the functions $l(t,x)$, $r(t,x)$ are constant along the characteristics

$$\begin{aligned} &\frac{dx_1}{dt} = -k(l-r)(t, x_1)\,, \quad x_1(0,\alpha) = \alpha,\\ &\frac{dx_2}{dt} = k(l-r)(t, x_2)\,, \quad x_2(0,\beta) = \beta, \end{aligned} \tag{6.9}$$

respectively, and that $l(t,x)$, $r(t,x)$ are periodic of period π with respect to $x \in \mathbb{R}$, i.e.,

$$l(t, x+\pi) = l(t,x)\,, \quad r(t, x+\pi) = r(t,x)\,. \tag{6.10}$$

Besides, we may assume that

$$l_0(0) = M\,. \tag{6.11}$$

Then, for $\varepsilon > 0$ sufficiently small, namely

$$0 < \varepsilon \leq \frac{M-m}{4}\,, \tag{6.12}$$

let us define with $D_\varepsilon = \big(\alpha_\varepsilon,\, \bar\alpha_\varepsilon\big)$ an open interval, containing 0, and such that

$$\begin{aligned} &l_0(x) \geq M - \varepsilon \quad \text{in} \quad D_\varepsilon,\\ &l_0(\alpha_\varepsilon)\,,\ l_0(\bar\alpha_\varepsilon) < M \quad \text{and} \quad l_0'(\bar\alpha_\varepsilon) < 0\,. \end{aligned} \tag{6.13}$$

Clearly, we have

$$\alpha_\varepsilon < 0 < \bar\alpha_\varepsilon \quad \text{and} \quad 0 < \bar\alpha_\varepsilon - \alpha_\varepsilon < \pi\,. \tag{6.14}$$

Besides, let us take an interval $I \subset \big(0,\, \pi\big)$ such that

$$\begin{aligned} &r_0(x) \leq \frac{m+M}{2}\,, \quad r_0'(x) \geq \rho \quad \text{for all } x \in I,\\ &\text{meas}\big(I\big) > 0 \end{aligned} \tag{6.15}$$

with $\rho > 0$. Finally, for $|\eta| \le M - m + 1$, let us define

$$\phi(\eta) \stackrel{\text{def}}{=} -\inf_{s \ge \eta} k'(s) \quad (|s| \le M - m + 1), \tag{6.16}$$

$\phi(\eta)$ is a decreasing function such that $\phi(\eta) > 0$ for $\eta < 0$ and $\phi(\eta) < 0$ for $\eta > 0$.

Then, we have the following.

LEMMA 2. *Let $\bar\alpha \in D_\varepsilon$ such that $l_0'(\bar\alpha) > 0$. Let $t(\beta) : \big(-\infty, \bar\alpha\big] \to \mathbb{R}$ be the C^1 function defined by the equation*

$$x_1\big(t, \bar\alpha\big) = x_2\big(t, \beta\big), \quad \text{with} \quad \beta \le \bar\alpha\,. \tag{6.17}$$

Then, for all $\beta < \bar\alpha$, we have

$$\frac{1}{\bar\alpha - \beta} \int_\beta^{\bar\alpha} \frac{\partial x_2}{\partial \beta}\big(t(y), y\big)\, \chi_{I+\pi\mathbb{Z}}(y)\, dy \le C \left(\frac{1}{l_0'(\bar\alpha)\,\big(\bar\alpha - \beta\big)} + \phi\big(l_0(\bar\alpha) - M\big) \right), \tag{6.18}$$

where $\chi_{I+\pi\mathbb{Z}}(\beta)$ is the characteristic function of the set $I + \pi\mathbb{Z}$ and the constant C is independent of $\bar\alpha \in D_\varepsilon$, $\beta \in \big(-\infty, \bar\alpha\big]$.

Proof. The condition $\frac{\partial x_1}{\partial \alpha}\big(t, \bar\alpha\big) > 0$ implies that (recalling (4.7))

$$\int_0^t k'\big(l_0(\bar\alpha) - r(\tau, x_1(\tau, \bar\alpha))\big)\, e^{h_1(0) - h_1(\tau)}\, d\tau < \frac{1}{l_0'(\bar\alpha)} \tag{6.19}$$

for all $t \ge 0$. Hence, using the change of variable $t = t(\beta)$,

$$\frac{dt}{d\beta} = \frac{-\dfrac{\partial x_2}{\partial \beta}\big(t(\beta), \beta\big)}{2\,k\big(l - r\big)\big(t(\beta), x_1(t(\beta), \bar\alpha)\big)}\,, \tag{6.20}$$

we obtain that

$$\int_\beta^{\bar\alpha} k'\big(l_0(\bar\alpha) - r_0(y)\big) \frac{\dfrac{\partial x_2}{\partial \beta}\big(t(y), y\big)\, e^{h_1(0) - h_1(t(y))}}{k\big(l - r\big)\big(t(y), x_1(t(y), \bar\alpha)\big)}\, dy < \frac{2}{l_0'(\bar\alpha)}\,. \tag{6.21}$$

Now, let us recall that by (2.15)

$$\frac{1}{C} \le \frac{e^{h_1(0) - h_1(t(\beta))}}{k\big(l - r\big)\big(t(\beta), x_1(t(\beta), \bar\alpha)\big)} \le C \tag{6.22}$$

for a suitable constant $C > 1$. Moreover, observe that

$$\beta \in I + \pi\mathbb{Z}\,, \quad \alpha \in D_\varepsilon \;\Rightarrow\; k'\big(l_0(\alpha) - r_0(\beta)\big) \ge \delta_1 \tag{6.23}$$

for a suitable $\delta_1 > 0$, and

$$\beta \notin I + \pi\mathbb{Z} \;\Rightarrow\; k'\big(l_0(\alpha) - r_0(\beta)\big) \ge -\,\phi\big(l_0(\alpha) - M\big), \tag{6.24}$$

where $\phi(\eta)$ is the function defined in (6.16).

Thus, taking into account of (6.22)–(6.24) we find

$$\frac{\delta_1}{C}\int_\beta^{\bar\alpha}\frac{\partial x_2}{\partial\beta}\big(t(y),y\big)\,\chi_{I+\pi\mathbb{Z}}(y)\,dy - C\phi\big(l_0(\bar\alpha)-M\big)\int_\beta^{\bar\alpha}\frac{\partial x_2}{\partial\beta}\big(t(y),y\big)\,dy < \frac{2}{l_0'(\bar\alpha)}\,. \tag{6.25}$$

Clearly, the estimate (6.18) follows immediately from (6.25) if we observe that

$$\frac{\bar\alpha-\beta}{C}\le\int_\beta^{\bar\alpha}\frac{\partial x_2}{\partial\beta}\big(t(y),y\big)\,dy\le C\,\big(\bar\alpha-\beta\big), \tag{6.26}$$

which is a consequence of the fact that, having

$$\frac{1}{\delta}\le k\big(l-r\big)(t,x)\le\delta \tag{6.27}$$

for some $\delta>1$, the function $t(\beta)$ satisfies

$$\frac{\bar\alpha-\beta}{C}\le t(\beta)\le C\big(\bar\alpha-\beta\big)\quad\text{for all }\beta\le\bar\alpha \tag{6.28}$$

for a suitable $C>1$. □

Remark 1. Observe that, introducing for $\omega>0$ the set

$$B_\omega=\Big\{\beta\le\bar\alpha\,\Big|\,\frac{\partial x_2}{\partial\beta}\big(t(\beta),\beta\big)\ge\omega,\quad\beta\in I+\pi\mathbb{Z}\Big\}, \tag{6.29}$$

then the estimate (6.18) gives for $\beta<\bar\alpha$

$$\frac{\operatorname{meas}\Big(B_\omega\cap[\beta,\,\bar\alpha]\Big)}{\bar\alpha-\beta}\le\frac{C}{\omega}\left(\frac{1}{l_0'(\bar\alpha)\big(\bar\alpha-\beta\big)}+\phi\big(l_0(\bar\alpha)-M\big)\right). \tag{6.30}$$

□

Now, let us consider the characteristics $x_2(t,\bar\beta)$ with $\bar\beta\le\bar\alpha$. From the results of section 2, we have easily

$$\begin{aligned}\frac{\partial x_2}{\partial\beta}\big(t,\bar\beta\big) = {}& e^{h_2(t)-h_2(t(\bar\beta))}\,\frac{\partial x_2}{\partial\beta}\big(t(\bar\beta),\bar\beta\big)\left(1-\frac{\partial r}{\partial x}\big(t(\bar\beta),x_2(t(\bar\beta),\bar\beta)\big)\right.\\ &\left.\cdot\int_{t(\bar\beta)}^{t}k'\big(l(\tau,x_2(\tau,\bar\beta))-r_0(\bar\beta)\big)\,e^{h_2(t(\bar\beta))-h_2(\tau)}\,d\tau\right),\end{aligned} \tag{6.31}$$

where

$$\frac{\partial r}{\partial x}\big(t(\bar\beta),x_2(t(\bar\beta),\bar\beta)\big)=\frac{r_0'(\bar\beta)}{\dfrac{\partial x_2}{\partial\beta}\big(t(\bar\beta),\bar\beta\big)}\,. \tag{6.32}$$

Hence, the condition $\frac{\partial x_2}{\partial\beta}\big(t(\bar\beta),\bar\beta\big)\le\omega$ for $\bar\beta\in I+\pi\mathbb{Z}$, $\bar\beta\notin B_\omega$ implies that

$$\frac{\partial r}{\partial x}\big(t(\bar\beta),x_2(t(\bar\beta),\bar\beta)\big)\ge\frac{\rho}{\omega}\,. \tag{6.33}$$

Moreover, observe that $l\big(t, x_2(t,\beta)\big) \geq M - \varepsilon$ gives

$$k'\big(l\big(t, x_2(t,\beta)\big) - r_0(\beta)\big) \geq \delta_1 > 0\,, \quad \text{if} \quad \beta \leq \bar\alpha,\ \beta \in I + \pi\mathbb{Z}\,, \tag{6.34}$$

with $\delta_1 > 0$ the constant of (6.23).

Thus, denoting with $t_\varepsilon(\beta)$ the unique solution of the equation

$$x_1\big(t,\, \bar\alpha_\varepsilon\big) = x_2(t,\, \beta) \quad \text{for} \quad \beta \leq \bar\alpha_\varepsilon\,, \tag{6.35}$$

from (6.31) and the condition $\frac{\partial x_2}{\partial \beta} > 0$, we easily have

$$\frac{\rho}{\omega}\ \delta_1\ \Big(t_\varepsilon(\bar\beta) - t(\bar\beta)\Big) \leq C \tag{6.36}$$

for all $\bar\beta \leq \bar\alpha$, $\bar\beta \in I + \pi\mathbb{Z}$, $\bar\beta \notin B_\omega$. Thus, we have proved the following.

LEMMA 3. *For all $\beta \leq \bar\alpha$, $\beta \in I + \pi\mathbb{Z}$, $\beta \notin B_\omega$ we have*

$$t_\varepsilon(\beta) - t(\beta)\ \leq\ C\,\omega, \tag{6.37}$$

where C does not depend on $\omega > 0$, ε, $\bar\alpha$, $\bar\alpha_\varepsilon$. □

To continue, let us fix $\bar\beta \leq \bar\alpha$, $\bar\beta \in I + \pi\mathbb{Z}$, $\bar\beta \notin B_\omega$ and let us evaluate the quantity

$$t_\varepsilon(\beta) - t(\beta) \tag{6.38}$$

for $\beta \leq \bar\alpha$ near $\bar\beta$. We have the following.

LEMMA 4. *Let $\bar\beta \leq \bar\alpha$, $\bar\beta \in I + \pi\mathbb{Z}$, $\bar\beta \notin B_\omega$, then for all $\beta \leq \bar\alpha$ we have*

$$t_\varepsilon(\beta) - t(\beta)\ \leq\ C\left(\omega + \big(\big|\beta - \bar\beta\big| + 1\big)\ \int_{\bar\alpha}^{\bar\alpha_\varepsilon} \big|l_0'(x)\big|\,dx\right), \tag{6.39}$$

where C is a constant independent of ω, β, $\bar\beta$, and $\bar\alpha \in D_\varepsilon$.

Proof. Let us observe that by (6.27) we have

$$\frac{1}{C}\big(t_\varepsilon(\beta) - t(\beta)\big) \leq x_1\big(t(\beta), \bar\alpha_\varepsilon\big) - x_1\big(t(\beta), \bar\alpha\big) \leq C\big(t_\varepsilon(\beta) - t(\beta)\big) \tag{6.40}$$

for a suitable constant $C > 1$. On the other hand, integrating $\frac{\partial x_1}{\partial \alpha}$ we find

(6.41)

$$\begin{aligned}
x_1\big(t(\beta), \bar\alpha_\varepsilon\big) - x_1\big(t(\beta), \bar\alpha\big) &= \int_{\bar\alpha}^{\bar\alpha_\varepsilon} e^{h_1(t(\beta),x) - h_1(0,x)} \\
&\cdot \left(1 - l_0'(x) \int_0^{t(\beta)} k'\big(l - r\big)(\tau, x_1(\tau,x))\, e^{h_1(0,x) - h_1(\tau,x)}\,d\tau\right) dx \\
&\leq C \int_{\bar\alpha}^{\bar\alpha_\varepsilon} e^{h_1(t(\bar\beta),x) - h_1(0,x)} \\
&\cdot \left(1 - l_0'(x) \int_0^{t(\beta)} k'\big(l - r\big)(\tau, x_1(\tau,x))\, e^{h_1(0,x) - h_1(\tau,x)}\,d\tau\right) dx \\
&\leq C\left(x_1\big(t(\bar\beta), \bar\alpha_\varepsilon\big) - x_1\big(t(\bar\beta), \bar\alpha\big)\right) \\
&+ C \int_{\bar\alpha}^{\bar\alpha_\varepsilon} \left(e^{h_1(t(\bar\beta),x) - h_1(0,x)}\, l_0'(x) \int_{t(\beta)}^{t(\bar\beta)} k'\big(l - r\big)(\tau, x_1(\tau,x))\, e^{h_1(0,x) - h_1(\tau,x)}\,d\tau\right) dx \\
&\leq C\,\omega + C\,\big|t(\beta) - t(\bar\beta)\big| \int_{\bar\alpha}^{\bar\alpha_\varepsilon} \big|l_0'(x)\big|\,dx\,.
\end{aligned}$$

Thus, to conclude the proof it is sufficient to observe that (thanks to the periodicity)

$$\big|t(\beta) - t(\bar\beta)\big| \,\le\, C\,\Big(\big|\beta-\bar\beta\big| + 1\Big) \tag{6.42}$$

for a suitable constant C. □

Remark 2. Let us observe that, since the functions $l(t,x)$, $r(t,x)$ are periodic of period $\pi > 0$ with respect to the variable x, it follows immediately that

$$t_\varepsilon(\beta) - t(\beta) \,\le\, C\,\pi \qquad \text{for all } \beta \le \bar\alpha\,, \tag{6.43}$$

because $\bar\alpha_\varepsilon - \bar\alpha \le \pi$. So the inequality (6.39) is a refinement of trivial estimate (6.43).

From (6.30) and the estimate of Lemma 4, we have the following.

LEMMA 5. *There exists a constant C_1 such that the inequality*

$$0 \le t_\varepsilon(\beta) - t(\beta) \le C_1\,\left(\omega + \int_{\bar\alpha}^{\bar\alpha_\varepsilon} \big|l_0'(x)\big|\,dx\right) \tag{6.44}$$

holds true for all $\beta \le \bar\alpha$ such that

$$\operatorname{dist}\big\{\beta,\, I + \pi\mathbb{Z} \setminus B_\omega\big\} \le 2\pi\,. \tag{6.45}$$

Denoting with $\tilde B_\omega$ the subset of $\beta \le \bar\alpha$ where (6.44) *does not hold, we have*

$$\frac{\operatorname{meas}\Big(\tilde B_\omega \,\cap\, \big[\beta,\,\bar\alpha\big]\Big)}{\bar\alpha - \beta} \,\le\, \frac{C}{\omega}\,\left(\frac{1}{l_0'(\bar\alpha)\big(\bar\alpha-\beta\big)} + \phi\big(l_0(\bar\alpha) - M\big)\right) \tag{6.46}$$

for all $\beta \le \bar\alpha - \pi$. The constants C_1, C are independent of β, $\bar\alpha \in D_\varepsilon$, and ω. □

Observation 1. If $x_0 \in \tilde B_\omega$ $(x_0 \le \bar\alpha - \pi)$, then there exists $k_0 \in \mathbb{Z}$ such that

$$k_0\pi \,\le\, x_0 \,\le\, (k_0+1)\pi \quad \text{ and } \quad I + (k_0-1)\pi\,,\; I + k_0\pi \subset B_\omega\,.$$

Let $L = \big\{k \in \mathbb{Z} \,:\, I + k\pi,\, I + (k+1)\pi \subset B_\omega\big\}$; then for all β, $\beta \le \bar\alpha - \pi$ we have

$$\operatorname{card}\big\{L \cap \big[\beta,\bar\alpha\big]\big\} \,\le\, C\,\operatorname{meas}\Big(B_\omega \cap \big[\beta,\bar\alpha\big]\Big)$$

and $\tilde B_\omega \subset \pi L + \big[0,\,2\pi\big]$. Note that (thanks to the periodicity) for all $t \ge 0$

$$\int_\beta^{\beta+\pi} \frac{\partial x_2}{\partial\beta}(t,y)\,dy = \pi \qquad \text{for all } \beta \in \mathbb{R}\,;$$

moreover, by (2.28), for a suitable constant $C \ge 1$ we have

$$\frac{\partial x_2}{\partial\beta}(\tau,\beta) \le C\,\frac{\partial x_2}{\partial\beta}(t,\beta) + C\,\big|\tau - t\big| \quad \text{for all} \quad t,\,\tau \ge 0\,.$$

This implies that, for $\beta \le \bar\alpha - \pi$,

$$\begin{aligned}\int_\beta^{\bar\alpha} \frac{\partial x_2}{\partial\beta}(t(y),y)\,\chi_{\tilde B_\omega}(y)\,dy \,&\le\, \int_\beta^{\bar\alpha} \frac{\partial x_2}{\partial\beta}(t(y),y)\,\chi_{\pi L + [0,2\pi]}(y)\,dy \\ &\le\, C\,\operatorname{meas}\Big(B_\omega \cap \big[\beta,\bar\alpha\big]\Big)\end{aligned} \tag{6.46a}$$

since $|y_1 - y_2| \le \pi \;\Rightarrow\; |t(y_1) - t(y_2)| \le C\pi$. □

To continue, let us introduce for $\gamma > 0$ the set

$$E_\gamma = \Big\{ \beta \in (0,\, \pi) \,\Big|\, r_0(\beta) \ge M - \gamma \Big\}. \tag{6.47}$$

Then, the following holds.

LEMMA 6. *Let us take $\bar\alpha \in D_\varepsilon$ such that $l_0'(\bar\alpha) > 0$, $M - \gamma < l_0(\bar\alpha)$. Then, for all $\beta < \bar\alpha$*

$$\begin{aligned} &\frac{1}{\bar\alpha - \beta} \int_\beta^{\bar\alpha} \frac{\partial x_2}{\partial \beta}\big(t(y), y\big)\, \chi_{E_\gamma + \pi\mathbb{Z}}(y)\, dy \\ &\qquad \ge \frac{1}{C_2} - \frac{C_2}{-\phi\big(l_0(\bar\alpha) - M + \gamma\big)} \left(\frac{1}{l_0'(\bar\alpha)\big(\bar\alpha - \beta\big)} + \phi\big(l_0(\bar\alpha) - M\big) \right), \end{aligned} \tag{6.48}$$

where $\chi_{E_\gamma + \pi\mathbb{Z}}(\beta)$ is the characteristic function of the set $E_\gamma + \pi\mathbb{Z}$; here $C_2 > 1$ is a suitable constant independent of $\bar\alpha$, β, γ.

Proof. To begin with let us recall that, by the inequalities (6.26),

$$\int_\beta^{\bar\alpha} \frac{\partial x_2}{\partial \beta}\big(t(y), y\big)\, dy = O\big(\bar\alpha - \beta\big).$$

Moreover, we have easily for $\beta \le \bar\alpha$

$$\beta \notin E_\gamma + \pi\mathbb{Z} \;\Rightarrow\; k'\big(l_0(\bar\alpha) - r_0(\beta)\big) \;\ge\; -\phi\big(l_0(\bar\alpha) - M + \gamma\big) > 0\,, \tag{6.49}$$

where $\phi(\eta)$ is defined in (6.16). Thus, from the condition $\frac{\partial x_1}{\partial \alpha}(t, \bar\alpha) > 0$ we find

$$\begin{aligned} &\frac{-\phi\big(l_0(\bar\alpha) - M + \gamma\big)}{C} \int_\beta^{\bar\alpha} \frac{\partial x_2}{\partial \beta}\big(t(y), y\big)\, \big(1 - \chi_{E_\gamma + \pi\mathbb{Z}}(y)\big)\, dy \\ &\qquad - C\phi\big(l_0(\bar\alpha) - M\big) \int_\beta^{\bar\alpha} \frac{\partial x_2}{\partial \beta}\big(t(y)\,, y\big)\, \chi_{E_\gamma + \pi\mathbb{Z}}(y)\, dy \;<\; \frac{2}{l_0'(\bar\alpha)}\,. \end{aligned} \tag{6.50}$$

Thus, we obtain that

$$\begin{aligned} &\int_\beta^{\bar\alpha} \frac{\partial x_2}{\partial \beta}\big(t(y), y\big)\, \big(1 - \chi_{E_\gamma + \pi\mathbb{Z}}(y)\big)\, dy \\ &\le \frac{C}{-\phi\big(l_0(\bar\alpha) - M + \gamma\big)} \left(\frac{1}{l_0'(\bar\alpha)} + \phi\big(l_0(\bar\alpha) - M\big) \int_\beta^{\bar\alpha} \frac{\partial x_2}{\partial \beta}\big(t(y), y\big)\, dy \right), \end{aligned} \tag{6.51}$$

which immediately gives (6.48). □

Remark 3. Let us remark that, since $l_0(x)$, $r_0(x)$ are nonconstant C^1 periodic functions of period $\pi > 0$ such that $\max l_0(x) = \max r_0(x) = M$, then we have

$$\begin{aligned} &0 < \operatorname{meas}\big(E_\gamma\big) \le \pi \quad \text{for all } \gamma > 0\,, \\ &\lim_{\varepsilon \to 0^+} \int_{\alpha_\varepsilon}^{\bar\alpha_\varepsilon} \big|l_0'(x)\big|\, dx = 0\,. \end{aligned} \tag{6.52}$$

The first equation of (6.52) is clear, while the other, recalling the definition of D_ε, is a simple exercise. □

We are now in position to conclude the proof of Case (ii) of Theorem 4. Let us evaluate

$$\frac{\partial x_1}{\partial \alpha}\big(t,\bar\alpha_\varepsilon\big) \quad \text{as} \quad t\to+\infty\,. \tag{6.53}$$

Taking into account that $l_0'(\bar\alpha_\varepsilon)<0$ we must have

$$\int_\beta^{\bar\alpha_\varepsilon} k'\big(l_0(\bar\alpha_\varepsilon)-r_0(y)\big)\frac{\dfrac{\partial x_2}{\partial\beta}\big(t_\varepsilon(y),y\big)\,e^{h_1(0)-h_1(t_\varepsilon(y))}}{k\big(l-r\big)\big(t_\varepsilon(y),x_1(t_\varepsilon(y),\bar\alpha_\varepsilon)\big)}\,dy \;>\; \frac{2}{l_0'(\bar\alpha_\varepsilon)} \tag{6.54}$$

for all $\beta\le\bar\alpha_\varepsilon$.

We will estimate the left-hand side of (6.54) by considering four cases:

(1) $y\le\bar\alpha$, $y\in E_\gamma+\pi\mathbb{Z}$, and $y\notin\tilde B_\omega$;
(2) $y\le\bar\alpha$ and $y\in E_\gamma+\pi\mathbb{Z}\cap\tilde B_\omega$;
(3) $y\le\bar\alpha$, $y\notin E_\gamma+\pi\mathbb{Z}$, and $y\notin\tilde B_\omega$;
(4) $y\le\bar\alpha$ and $y\notin E_\gamma+\pi\mathbb{Z}$, $y\in\tilde B_\omega$.

Here $\tilde B_\omega$ is the set introduced in Lemma 5. Denoting the four sets defined above by J_1, J_2, J_3, and J_4, respectively, and assuming that

$$\begin{aligned}&0<\gamma\le\frac{M-l_0(\bar\alpha_\varepsilon)}{4}\,,\\ &\bar\alpha\in D_\varepsilon\,,\quad l_0'\big(\bar\alpha\big)>0\,,\quad M-\gamma<l_0\big(\bar\alpha\big)\,,\end{aligned}$$

according to the statement of Lemma 6, we have the following.

(1) $y\in J_1\;\Rightarrow\; l_0(\bar\alpha_\varepsilon)-M\le l_0(\bar\alpha_\varepsilon)-r_0(y)\le l_0(\bar\alpha_\varepsilon)-M+\gamma$, thus

$$k'\big(l_0(\bar\alpha_\varepsilon)-r_0(y)\big)\le-\psi\big(l_0(\bar\alpha_\varepsilon),\gamma\big), \tag{6.55}$$

where $\psi\big(l_0(\bar\alpha_\varepsilon),\gamma\big)$ is defined by

$$\psi\big(l_0(\bar\alpha_\varepsilon),\gamma\big)\;\stackrel{\text{def}}{=}\;-\sup\,k'(s)\quad\text{for}\quad l_0(\bar\alpha_\varepsilon)-M\le s\le l_0(\bar\alpha_\varepsilon)-M+\gamma\,. \tag{6.56}$$

Hence, we have

$$\int_\beta^{\bar\alpha}\big\{\,*\,\big\}\,\chi_{J_1}(y)\,dy\le-C\,\psi\big(l_0(\bar\alpha_\varepsilon),\gamma\big)\int_\beta^{\bar\alpha}\frac{\partial x_2}{\partial\beta}\big(t_\varepsilon(y),y\big)\,\chi_{J_1}(y)\,dy\,, \tag{6.57}$$

where $\big\{\,*\,\big\}$ represents the function in the integral of (6.54).

Now, recalling (2.28) of section 2, let us observe that

$$\frac{\partial x_2}{\partial\beta}\big(t_\varepsilon(y),y\big)\ge\frac{1}{C}\frac{\partial x_2}{\partial\beta}\big(t(y),y\big)-C\big|t_\varepsilon(y)-t(y)\big| \tag{6.58}$$

for a suitable constant $C>1$. Thus, we may conclude that

$$\begin{aligned}\int_\beta^{\bar\alpha}\big\{\,*\,\big\}\,\chi_{J_1}(y)\,dy\le&-\frac{1}{C}\,\psi\big(l_0(\bar\alpha_\varepsilon),\gamma\big)\int_\beta^{\bar\alpha}\frac{\partial x_2}{\partial\beta}\big(t(y),y\big)\,\chi_{J_1}(y)\,dy\\ &+C\,\psi\big(l_0(\bar\alpha_\varepsilon),\gamma\big)\max_{y\notin\tilde B_\omega}\big|t_\varepsilon(y)-t(y)\big|\int_\beta^{\bar\alpha}\chi_{J_1}(y)\,dy,\end{aligned} \tag{6.59}$$

where $\max_{y\notin\tilde B_\omega}\big|t_\varepsilon(y)-t(y)\big|$ may be small, according to (6.44) of Lemma 5. □

(2) By the same argument of case (1), we know that $k'\big(l_0(\bar\alpha_\varepsilon)-r_0(y)\big)\le -\psi\big(l_0(\bar\alpha_\varepsilon),\gamma\big)$. But now, since $y\in\tilde B_\omega$, we have only (recall Remark 2)

$$\big|t_\varepsilon(y)-t(y)\big|\le C\pi\,. \tag{6.60}$$

Thus, we find

$$\begin{aligned}\int_\beta^{\bar\alpha}\big\{\,*\,\big\}\,\chi_{J_2}(y)\,dy \le -\frac{1}{C}\,\psi\big(l_0(\bar\alpha_\varepsilon),\gamma\big)\int_\beta^{\bar\alpha}\frac{\partial x_2}{\partial\beta}\big(t(y),y\big)\,\chi_{J_2}(y)\,dy\\ +C\,\pi\,\psi\big(l_0(\bar\alpha_\varepsilon),\gamma\big)\int_\beta^{\bar\alpha}\chi_{J_2}(y)\,dy\,.\qquad\square\end{aligned} \tag{6.61}$$

(3) For $y\notin E_\gamma+\pi\mathbb{Z}$ we have only

$$\big|k'\big(l_0(\bar\alpha_\varepsilon)-r_0(y)\big)\big|\le C\,. \tag{6.62}$$

In this case, to estimate $\int_\beta^{\bar\alpha}\big\{\,*\,\big\}\,\chi_{J_3}(y)\,dy$, let us begin by observing that

$$t_\varepsilon(\beta)=-\int_\beta^{\bar\alpha_\varepsilon}\frac{dt_\varepsilon}{d\beta}\,dy\,,\quad t(\beta)=-\int_\beta^{\bar\alpha}\frac{dt}{d\beta}\,dy \tag{6.63}$$

and that from (6.43), having $0\le\bar\alpha_\varepsilon-\bar\alpha\le\pi$,

$$0\le t_\varepsilon(\beta)-t(\beta)\le C\,\pi,\quad \left|\int_{\bar\alpha}^{\bar\alpha_\varepsilon}\frac{dt_\varepsilon}{d\beta}\,dy\right|\le C\,\pi\,. \tag{6.64}$$

Then, thanks to (6.20), the inequalities of (6.64) imply the following:

$$\left|\int_\beta^{\bar\alpha}\frac{\dfrac{\partial x_2}{\partial\beta}\big(t_\varepsilon(y),y\big)\,dy}{k(l-r)\big(t_\varepsilon(y),x_1(t_\varepsilon(y),\bar\alpha_\varepsilon)\big)}-\int_\beta^{\bar\alpha}\frac{\dfrac{\partial x_2}{\partial\beta}\big(t(y),y\big)\,dy}{k(l-r)\big(t(y),x_1(t(y),\bar\alpha)\big)}\right|\le C\,\pi\,. \tag{6.65}$$

Moreover, from (2.11) and (2.12) of section 2 we have the relations

$$\frac{\partial x_2}{\partial\beta}(t,y)=e^{h_2(t)-h_2(0)}\left\{1-r_0'(y)\int_0^t k'\big(l-r\big)\big(\tau,x_2(\tau,y)\big)\,e^{h_2(0)-h_2(\tau)}\,d\tau\right\} \tag{6.66}$$

with

$$h_2(t)-h_2(0)=\frac{1}{2}\ln\Big(k\big(l-r\big)\big(t,x_2(t,y)\big)\Big)-\frac{1}{2}\ln\Big(k\big(l_0(y)-r_0(y)\big)\Big)\,. \tag{6.67}$$

Thus, we easily find that

$$\begin{aligned}\frac{\dfrac{\partial x_2}{\partial\beta}\big(t_\varepsilon(y),y\big)}{k(l-r)\big(t_\varepsilon(y),x_1(t_\varepsilon(y),\bar\alpha_\varepsilon)\big)}&=\frac{k(l-r)\big(t(y),x_1(t(y),\bar\alpha)\big)^{1/2}}{k(l-r)\big(t_\varepsilon(y),x_1(t_\varepsilon(y),\bar\alpha_\varepsilon)\big)^{1/2}}\\ &\quad\cdot\frac{\dfrac{\partial x_2}{\partial\beta}\big(t(y),y\big)}{k(l-r)\big(t(y),x_1(t(y),\bar\alpha)\big)}\\ &\quad-\frac{r_0'(y)\displaystyle\int_{t(y)}^{t_\varepsilon(y)}k'(l-r)\big(\tau,x_2(\tau,y)\big)\,e^{h_2(0)-h_2(\tau)}\,d\tau}{k(l-r)\big(t_\varepsilon(y),x_1(t_\varepsilon(y),\bar\alpha_\varepsilon)\big)^{1/2}k\big(l_0(y)-r_0(y)\big)^{1/2}}\,,\end{aligned} \tag{6.68}$$

where

$$\frac{k\big(l-r\big)\big(t(y),x_1(t(y),\bar\alpha)\big)^{1/2}}{k\big(l-r\big)\big(t_\varepsilon(y),x_1(t_\varepsilon(y),\bar\alpha_\varepsilon)\big)^{1/2}}=\frac{k\big(l_0(\bar\alpha)-r_0(y)\big)^{1/2}}{k\big(l_0(\bar\alpha_\varepsilon)-r_0(y)\big)^{1/2}}\,. \tag{6.69}$$

Moreover, we can rewrite the above quotient as

$$\begin{aligned}&\frac{k\big(l_0(\bar\alpha)-r_0(y)\big)^{1/2}}{k\big(l_0(\bar\alpha_\varepsilon)-r_0(y)\big)^{1/2}}\\ &=1-\frac{k\big(l_0(\bar\alpha_\varepsilon)-r_0(y)\big)-k\big(l_0(\bar\alpha)-r_0(y)\big)}{k\big(l_0(\bar\alpha_\varepsilon)-r_0(y)\big)^{1/2}\big[k\big(l_0(\bar\alpha)-r_0(y)\big)^{1/2}+k\big(l_0(\bar\alpha_\varepsilon)-r_0(y)\big)^{1/2}\big]}\\ &\stackrel{\text{def}}{=}1-Q\big(y,\bar\alpha,\bar\alpha_\varepsilon\big)\,.\end{aligned} \tag{6.70}$$

Thus, integrating over the set J_1, we find

$$\begin{aligned}\int_\beta^{\bar\alpha}\frac{\dfrac{\partial x_2}{\partial\beta}\big(t_\varepsilon(y),y\big)\,\chi_{J_1}(y)\,dy}{k(l-r)\big(t_\varepsilon(y),x_1(t_\varepsilon(y),\bar\alpha_\varepsilon)\big)}&=\int_\beta^{\bar\alpha}\frac{\dfrac{\partial x_2}{\partial\beta}\big(t(y),y\big)\,\chi_{J_1}(y)\,dy}{k(l-r)\big(t(y),x_1(t(y),\bar\alpha)\big)}\\ &-\int_\beta^{\bar\alpha}\frac{\dfrac{\partial x_2}{\partial\beta}\big(t(y),y\big)}{k(l-r)\big(t(y),x_1(t(y),\bar\alpha)\big)}\,Q\big(y,\bar\alpha,\bar\alpha_\varepsilon\big)\,\chi_{J_1}(y)\,dy\\ &-\int_\beta^{\bar\alpha}\frac{r_0'(y)\displaystyle\int_{t(y)}^{t_\varepsilon(y)}k'(l-r)\big(\tau,x_2(\tau,y)\big)\,e^{h_2(0)-h_2(\tau)}\,d\tau}{k(l-r)\big(t_\varepsilon(y),x_1(t_\varepsilon(y),\bar\alpha_\varepsilon)\big)^{1/2}k\big(l_0(y)-r_0(y)\big)^{1/2}}\,\chi_{J_1}(y)\,dy\,.\end{aligned} \tag{6.71}$$

Now, by the definition of J_1 and from the estimates (6.30), (6.46a), and (6.51) (recall the definition of $\tilde B_\omega$ and observation (1), for $\beta\le\bar\alpha-\pi$ we have

$$\int_\beta^{\bar\alpha}\frac{\dfrac{\partial x_2}{\partial\beta}\big(t(y),y\big)}{2\,k(l-r)\big(t(y),x_1(t(y),\bar\alpha)\big)}\,\chi_{J_1}(y)\,dy\ge t(\beta)-C\,\big(\bar\alpha-\beta\big)\,Y\big(\gamma,\omega,\bar\alpha,\beta\big), \tag{6.72}$$

where the quantity $Y=Y\big(\gamma,\omega,\bar\alpha,\beta\big)$ is given by

$$Y\stackrel{\text{def}}{=}\left(\frac{1}{\omega}+\frac{1}{-\phi\big(l_0(\bar\alpha)-M+\gamma\big)}\right)\left(\frac{1}{l_0'(\bar\alpha)\big(\bar\alpha-\beta\big)}+\phi\big(l_0(\bar\alpha)-M\big)\right). \tag{6.73}$$

Hence, from (6.71), (6.72), and (6.26) we easily have

$$\begin{aligned}\int_\beta^{\bar\alpha}\frac{\dfrac{\partial x_2}{\partial\beta}\big(t_\varepsilon(y),y\big)\,\chi_{J_1}(y)dy}{2\,k(l-r)\big(t_\varepsilon(y),x_1(t_\varepsilon(y),\bar\alpha_\varepsilon)\big)}\ &\ge\ t(\beta)-C\big(\bar\alpha-\beta\big)\,Y\big(\gamma,\omega,\bar\alpha,\beta\big)\\ &-C\,\big(\bar\alpha-\beta\big)\sup_{y\in J_1}Q\big(y,\bar\alpha,\bar\alpha_\varepsilon\big)\\ &-C\big(\bar\alpha-\beta\big)\sup_{y\in J_1}\big|r_0'(y)\big|\,\cdot\,\sup_{y\in J_1}\big|t_\varepsilon(y)-t(y)\big|\,.\end{aligned} \tag{6.74}$$

Now, taking into account that $\frac{\partial x_2}{\partial \beta}(t,y) > 0$ and that by (6.65)

$$\int_\beta^{\bar\alpha} \frac{\dfrac{\partial x_2}{\partial \beta}\big(t_\varepsilon(y),y\big)\,dy}{2\,k(l-r)\big(t_\varepsilon(y),x_1(t_\varepsilon(y),\bar\alpha_\varepsilon)\big)} \;\le\; t(\beta) + C\,\pi\,,$$

from (6.74) we find for the integral over J_3:

$$\begin{aligned} 0 \le \int_\beta^{\bar\alpha} \frac{\dfrac{\partial x_2}{\partial \beta}\big(t_\varepsilon(y),y\big)\,\chi_{J_3}(y)dy}{k(l-r)\big(t_\varepsilon(y),x_1(t_\varepsilon(y),\bar\alpha_\varepsilon)\big)} &\le C\,\pi + C\big(\bar\alpha-\beta\big)\,Y\big(\gamma,\omega,\bar\alpha,\beta\big) \\ &+ C\,\big(\bar\alpha-\beta\big)\,\sup_{y\in J_1} Q\big(y,\bar\alpha,\bar\alpha_\varepsilon\big) \\ &+ C\big(\bar\alpha-\beta\big)\,\sup_{y\in J_1}\big|r_0'(y)\big|\,\cdot\,\sup_{y\in J_1}\big|t_\varepsilon(y)-t(y)\big|\,. \end{aligned} \tag{6.75}$$

Then, recalling (6.62) and that in inequality (6.54) the term $e^{h_1(0)-h_1(t_\varepsilon(y))}$ is uniformly bounded, we immediately obtain the estimate

$$\begin{aligned} \int_\beta^{\bar\alpha} \big\{\,*\,\big\}\chi_{J_3}(y)\,dy \le C\,\pi + C\big(\bar\alpha-\beta\big)\,Y\big(\gamma,\omega,\bar\alpha,\beta\big) \\ +\; C\,\big(\bar\alpha-\beta\big)\,\sup_{y\in J_1} Q\big(y,\bar\alpha,\bar\alpha_\varepsilon\big) \\ +\; C\,\big(\bar\alpha-\beta\big)\,\sup_{y\in J_1}\big|r_0'(y)\big|\,\cdot\,\sup_{y\in J_1}\big|t_\varepsilon(y)-t(y)\big|\,. \end{aligned} \tag{6.76}$$

Remark 4. In the following we will always assume that $0<\gamma\le\frac{M-l_0(\bar\alpha_\varepsilon)}{4}$ and that $l_0(\bar\alpha) > M-\gamma$. Thus, for $y\in J_1\subset E_\gamma+\pi\mathbb{Z}$, we have

$$l_0(\bar\alpha_\varepsilon) \;<\; M-\gamma \;\le\; r_0(y)\,,\; l_0(\bar\alpha) \;\le M\,. \tag{6.77}$$

Moreover, let us observe that $k(\eta)$ is strictly decreasing for $\eta<0$ and strictly increasing for $\eta>0$; in fact, by (1.12),

$$k'(\eta)<0 \quad \text{for} \quad \eta<0\,, \quad k'(\eta)>0 \quad \text{for} \quad \eta>0\,.$$

Thus, for $\gamma>0$ sufficiently small from (6.27), (6.70) it follows that $Q\big(y,\bar\alpha,\bar\alpha_\varepsilon\big)\ge 0$ and

$$\begin{aligned} \sup_{y\in J_1} Q\big(y,\bar\alpha,\bar\alpha_\varepsilon\big) \;&\le \frac{\delta}{2}\,\sup_{y\in J_1}\Big(k\big(l_0(\bar\alpha_\varepsilon)-r_0(y)\big)-k(0)\Big) \\ &\le \frac{\delta}{2}\,\sup_{l_0(\bar\alpha_\varepsilon)-M\le\eta\le 0}\Big(k(\eta)-k(0)\Big) \\ &= \frac{\delta}{2}\,\Big(k\big(l_0(\bar\alpha_\varepsilon)-M\big)-k(0)\Big)\,. \end{aligned} \tag{6.78}$$

The last estimate will be important in the following. □

(4) Finally, if $y\notin E_\gamma+\pi\mathbb{Z}$, $y\in\tilde B_\omega$, we easily obtain

$$\begin{aligned} \int_\beta^{\bar\alpha}\big\{\,*\,\big\}\,\chi_{J_4}(y)\,dy \le C\int_\beta^{\bar\alpha}\frac{\partial x_2}{\partial\beta}\big(t(y),y\big)\,\chi_{J_4}(y)\,dy \\ +\,C\,\pi\int_\beta^{\bar\alpha}\chi_{J_4}(y)\,dy\,. \end{aligned} \tag{6.79}$$

□

Remark 5. Let us remark that the constants appearing in the estimates (6.59), (6.61), (6.76), (6.79) are independent of ε, γ, $\omega > 0$, $\bar\alpha \in D_\varepsilon$, and $\beta < \bar\alpha$. Thus, taking $C_0 > 1$ sufficiently large, in the following we will replace all the above constants by C_0. □

Putting together the above estimates, we have the following.

LEMMA 7. *Let us choose* ε, $D_\varepsilon = \big(\alpha_\varepsilon, \bar\alpha_\varepsilon\big)$ *according to* (6.12)–(6.14) *and* $\omega > 0$ *such that*

$$\left(\omega + \int_{\alpha_\varepsilon}^{\bar\alpha_\varepsilon} \big|l_0'(x)\big|\,dx\right) \le \frac{1}{32}\,\frac{1}{C_0^2\,C_1\,C_2}, \tag{6.80}$$

where C_1, C_2 *are the constants appearing, respectively, in* (6.44) *and* (6.48).

Moreover, assume that

$$\delta\left(k\big(l_0(\bar\alpha_\varepsilon) - M\big) - k(0)\right) \le -\frac{1}{32}\,\frac{1}{C_0^2\,C_2}\,k'\big(l_0(\bar\alpha_\varepsilon) - M\big)\,, \tag{6.81}$$

where $\delta > 1$ *is the constant in* (6.27). *Then, taking* $\gamma \le \big(M - l_0(\bar\alpha_\varepsilon)\big)/4$ *and* $\bar\alpha \in D_\varepsilon$ *such that*

$$l_0'(\bar\alpha) > 0 \quad \textit{and} \quad l_0(\bar\alpha) > M - \gamma\,, \tag{6.82}$$

with $\gamma > 0$ *sufficiently small and* $l_0(\bar\alpha)$ *sufficiently close to* M, *the above estimates imply that in* (6.54)

$$\int_\beta^{\bar\alpha} \big\{\,*\,\big\}\,dy \,\to\, -\infty, \quad \textit{as} \quad \beta \,\to\, -\infty\,. \tag{6.83}$$

Noting that $\big|\int_{\bar\alpha}^{\bar\alpha_\varepsilon}\{*\}\,dy\big| \le C$, (6.83) *proves the formation of singularities in finite time. Hence, if* (6.80), (6.81) *hold, we have proved Case* (ii) *of Theorem* 4.

Proof. From the estimates (6.59), (6.61), (6.76), (6.79) (recall Remark 5) we have

$$\begin{aligned}
\int_\beta^{\bar\alpha} \big\{\,*\,\big\}\,dy \le{}& -\frac{1}{C_0}\,\psi\big(l_0(\bar\alpha_\varepsilon),\gamma\big)\int_\beta^{\bar\alpha}\frac{\partial x_2}{\partial\beta}\big(t(y),y\big)\,\chi_{J_1\cup J_2}(y)\,dy \\
&+ C_0\,\psi\big(l_0(\bar\alpha_\varepsilon),\gamma\big)\,\max_{y\notin\tilde B_\omega}\big|t_\varepsilon(y) - t(y)\big|\int_\beta^{\bar\alpha}\chi_{J_1}(y)\,dy \\
&+ C_0\,\pi\,\psi\big(l_0(\bar\alpha_\varepsilon),\gamma\big)\int_\beta^{\bar\alpha}\chi_{J_2}(y)\,dy \\
&+ C_0\int_\beta^{\bar\alpha}\frac{\partial x_2}{\partial\beta}\big(t(y),y\big)\,\chi_{J_4}(y)\,dy \\
&+ C_0\,\pi\int_\beta^{\bar\alpha}\chi_{J_4}(y)\,dy \\
&+ \int_\beta^{\bar\alpha}\big\{\,*\,\big\}\,\chi_{J_3}(y)\,dy,
\end{aligned} \tag{6.84}$$

where from the definition (6.56) we know that $\psi\big(l_0(\bar\alpha_\varepsilon),\gamma\big)$ is strictly positive because $\gamma \le \big(M - l_0(\bar\alpha_\varepsilon)\big)/4$.

Let us analyze the terms of (6.84). First, from (6.48) we have

$$\frac{1}{C_0}\,\psi\big(l_0(\bar\alpha_\varepsilon),\gamma\big)\int_\beta^{\bar\alpha}\frac{\partial x_2}{\partial\beta}\big(t(y),y\big)\,\chi_{J_1\cup J_2}(y)\,dy \geq \frac{1}{C_0}\,\psi\big(l_0(\bar\alpha_\varepsilon),\gamma\big)\,\frac{\bar\alpha-\beta}{2\,C_2} \tag{6.85}$$
$$\stackrel{\text{def}}{=} K_\gamma\,\big(\bar\alpha-\beta\big),$$

provided $l_0'\big(\bar\alpha\big)>0$ and

$$l_0(\bar\alpha)>M-\gamma\,,\quad l_0(\bar\alpha)\quad\text{is sufficiently close to}\quad M,\quad\text{and}\quad \beta\ll\bar\alpha\ . \tag{6.86}$$

Remark 6. Let us remark again that, by (6.56), $\psi\big(l_0(\bar\alpha_\varepsilon),\gamma\big)>0$ for $0<\gamma\leq\big(M-l_0(\bar\alpha_\varepsilon)\big)/4$; moreover, since $k(\eta)\in C^1$, we have

$$\lim_{\gamma\to0^+}\psi\big(l_0(\bar\alpha_\varepsilon),\gamma\big)=-k'\big(l_0(\bar\alpha_\varepsilon)-M\big)>0\,. \tag{6.87}$$

Thus, we may assume that

$$K\ \leq\ K_\gamma\ \leq\ 2\,K \tag{6.88}$$

for a suitable $K>0$. This fact permits us to estimate the terms of (6.84) *uniformly* with respect to γ, for $\gamma>0$ sufficiently small. □

Next, thanks to (6.46), it follows that

$$C_0\,\pi\,\Big(1+\psi\big(l_0(\bar\alpha_\varepsilon),\gamma\big)\Big)\int_\beta^{\bar\alpha}\chi_{J_2\cup J_4}(y)\,dy\leq\frac{K}{8}\,\big(\bar\alpha-\beta\big), \tag{6.89}$$

provided again that $l_0(\bar\alpha)$ is sufficiently close to M and $\beta\ll\bar\alpha$.

From the estimate (6.51) of the proof of Lemma 6, we find also that

$$C_0\int_\beta^{\bar\alpha}\frac{\partial x_2}{\partial\beta}\big(t(y),y\big)\,\chi_{J_4}(y)\,dy\leq\frac{K}{8}\,\big(\bar\alpha-\beta\big), \tag{6.90}$$

provided the conditions (6.86) are satisfied.

Now, let us consider the second term in (6.84). Using (6.44) and the condition (6.80), we may estimate

$$t_\varepsilon(y)-t(y)\quad\text{for}\quad y\notin\tilde B_\omega\,;$$

then, from (6.88) and the definition of K_γ, we obtain that

$$C_0\,\psi\big(l_0(\bar\alpha_\varepsilon),\gamma\big)\,\max_{y\notin\tilde B_\omega}\big|t_\varepsilon(y)-t(y)\big|\int_\beta^{\bar\alpha}\chi_{J_1}(y)\,dy\leq\frac{K}{8}\,\big(\bar\alpha-\beta\big)\,. \tag{6.91}$$

Finally, let us estimate the term $\int_\beta^{\bar\alpha}\big\{*\big\}\,\chi_{J_3}(y)\,dy$. From (6.43), (6.76), and the result of Remark 4, for $\gamma>0$ sufficiently small, we may write

$$\begin{aligned}\int_\beta^{\bar\alpha}\big\{\,*\,\big\}\chi_{J_3}(y)\,dy\leq{}& C_0\,\pi+C_0\,\big(\bar\alpha-\beta\big)\,Y\big(\gamma,\omega,\bar\alpha,\beta\big)\\ &+C_0\,\frac{\delta}{2}\,\big(\bar\alpha-\beta\big)\,\Big(k\big(l_0(\bar\alpha_\varepsilon)-M\big)-k(0)\Big)\\ &+C_0\big(\bar\alpha-\beta\big)\sup_{y\in J_1}\big|r_0'(y)\big|\ C\,\pi\,,\end{aligned} \tag{6.92}$$

where we have easily

$$C_0\, Y\big(\gamma,\omega,\bar\alpha,\beta\big) \le \frac{K}{8}\,, \quad C_0 \sup_{y\in J_1} \big|r_0'(y)\big|\, C\,\pi \le \frac{K}{8}\,,$$

provided $\gamma > 0$ is sufficiently small and (6.86) holds (note that $\sup_{E_\gamma} |r_0'(y)| \to 0$ as $\gamma \to 0^+$).

Moreover, by assumption (6.81) on $l_0(\bar\alpha_\varepsilon)$ and definition (6.56) of $\psi\big(l_0(\bar\alpha_\varepsilon),\gamma\big)$ we have also the inequality

$$\frac{\delta}{2}\, C_0 \Big(k\big(l_0(\bar\alpha_\varepsilon) - M\big) - k(0)\Big) \le \frac{1}{32}\frac{1}{C_0\, C_2}\,\psi\big(l_0(\bar\alpha_\varepsilon),\gamma\big) \;\le\; \frac{K}{8} \tag{6.93}$$

for $\gamma > 0$ sufficiently small.

Thus, summing up the above estimates, we have proved that

$$\int_\beta^{\bar\alpha} \big\{\, *\,\big\}\, dy \le C_0\pi - \frac{K}{8}\,\big(\bar\alpha - \beta\big), \tag{6.94}$$

provided $\gamma > 0$ is sufficiently small and (6.86) holds. This completes the proof of Lemma 7. □

Conclusion of the proof of case (ii) of Theorem 4. To complete the proof of case (ii) it is now sufficient to find $D_\varepsilon = \big(\alpha_\varepsilon,\, \bar\alpha_\varepsilon\big)$, $\omega > 0$ such that the conditions (6.12)–(6.14) and (6.80), (6.81) are satisfied.

First, let us observe that, since $k(\eta) \in C^1$ and

$$k'(\eta) < 0 \quad \text{for} \quad \eta < 0\,,$$

for every $R > 0$ we can find a sequence $\big\{\eta_n\big\}$ such that $\eta_n < 0\,,\ \eta_n \to 0$, and

$$k'\big(\eta_n\big) \le -\,R\,\Big(k\big(\eta_n\big) - k(0)\Big)\,. \tag{6.95}$$

Thus, for fixed

$$R = 32\, C_0^2\, C_2\, \delta\,,$$

and taking into account that $\max l_0(x) = l_0(0) = M$, we can find easily a sequence $\big\{\alpha_n\big\}\,,\ 0 < \alpha_n < \pi$, such that, as $n \to \infty$,

$$l_0\big(\alpha_n\big) \to M\,, \quad l_0'(\alpha_n) < 0 \tag{6.96}$$

(thanks to Sard's theorem) and

$$k'\big(l_0\big(\alpha_n\big) - M\big) \;\le\; -32\, C_0^2\, C_2\, \delta\,\Big(k\big(l_0(\alpha_n) - M\big) - k(0)\Big)\,. \tag{6.97}$$

Hence, we have verified the condition (6.81).

Moreover, let us observe that for any fixed $n \in \mathbb{N}$, for $\gamma > 0$ sufficiently small, we have also

$$\sup\, k'(s) \;\le\; -16\, C_0^2\, C_2\, \delta\,\Big(k\big(l_0(\alpha_n) - M\big) - k(0)\Big) \tag{6.98}$$

for $l_0\big(\alpha_n\big) - M \le s \le l_0\big(\alpha_n\big) - M + \gamma\,.$

Thus, it is clear that for any fixed $n \in \mathbb{N}$ we have

$$\frac{\delta}{2}\Big(k\big(l_0(\alpha_n) - M\big) - k(0)\Big) \le \frac{1}{32}\frac{1}{C_0^2\, C_2}\,\psi\big(l_0(\alpha_n), \gamma\big) \le \frac{K}{8} \tag{6.99}$$

(that is, (6.93)), provided $\gamma > 0$ is sufficiently small.

To verify also (6.12)–(6.14) and (6.80) (for $\omega > 0$ small), we may easily choose the sequence $\{\alpha_n\}$ which satisfies (6.96), such that

$$\inf_{[0,\alpha_n]} l_0(x) \;\to\; M \tag{6.100}$$

as $n \to \infty$. Clearly, this implies that

$$\lim_{n\to\infty} \int_0^{\alpha_n} \big|l_0'(x)\big|\, dx = 0\,. \tag{6.101}$$

Then, taking for $n_0 \in \mathbb{N}$ sufficiently large

$$\bar{\alpha}_\varepsilon = \alpha_{n_0} \tag{6.102}$$

and $\alpha_\varepsilon < 0$ sufficiently close to 0 all the conditions of Lemma 7 are satisfied. □

REFERENCES

[1] P.H. Chang, *On the breakdown phenomena of solutions of quasi-linear wave equations*, Michigan Math. J., 23 (1976), pp. 277–287.

[2] F. Colombini and D. Del Santo, *Formation of singularities for nonlinear hyperbolic 2×2 systems with periodic data*, Osaka J. Math., 34 (1997), pp. 99–113.

[3] A. Douglis, *Some existence theorems for hyperbolic systems of partial differential equations in two independent variables*, Comm. Pure Appl. Math., 2 (1952), pp. 119–154.

[4] J. Glimm and P.D. Lax, *Decay of Solutions of Systems of Nonlinear Hyperbolic Conservation Laws*, Mem. Amer. Math. Soc. 101, AMS, Providence, RI, 1970.

[5] P. Hartman and A. Wintner, *Hyperbolic partial differential equations*, Amer. J. Math., 74 (1952), pp. 834–864.

[6] F. John, *Formation of singularities in one dimensional nonlinear wave propagation*, Comm. Pure Appl. Math., 27 (1974), pp. 377–405.

[7] F. John, *Nonlinear Wave Equations, Formations of Singularities*, Univ. Lecture Ser. 2, AMS, Providence, RI, 1990.

[8] J.L. Johnson, *Global continuous solutions of hyperbolic systems of quasi-linear equations*, Bull. Amer. Math. Soc., 73 (1967), pp. 639–641.

[9] J.B. Keller and Lu Ting, *Periodic vibrations of systems governed by nonlinear partial differential equations*, Comm. Pure Appl. Math., 19 (1966), pp. 371–420.

[10] S. Klainerman and A. Majda, *Formation of singularities for wave equations including the nonlinear vibrating string*, Comm. Pure Appl. Math., 33 (1980), pp. 241–263.

[11] Kuo-Shung Cheng, *Formation of singularities for nonlinear hyperbolic partial differential equations*, Contemp. Math., 17 (1983), pp. 45–56.

[12] Li Ta-tsien and Kong De-xing, *Blow up of periodic solutions to quasi-linear hyperbolic systems*, Nonlinear Anal., 26 (1996), pp. 1779–1789.

[13] Li Ta-tsien, Zhou Yi, and Kong De-xing, *Weak linear degeneracy and global classical solutions for general quasi-linear hyperbolic systems*, Comm. Partial Differential Equations, 19 (1994), pp. 1263–1317.

[14] Li Ta-tsien, *Global regularity and formation of singularities of solutions to first order quasilinear hyperbolic systems*, Proc. Royal Soc. Edinburgh Sect. A, 87 (1981), pp. 255–261.

[15] R.C. MacCamy and V.J. Mizel, *Existence and nonexistence in the large of solutions of quasilinear wave equations*, Arch. Rational Mech. Anal., 25 (1967), pp. 299–320.

[16] A. Majda, *Compressible Fluid Flow and Systems of Conservations Laws in Several Variables*, Springer-Verlag, New York, 1984.

[17] Tai-Ping Liu, *Development of singularities in the nonlinear waves for quasi-linear hyperbolic partial differential equations*, J. Differential Equations, 33 (1979), pp. 92–111.

[18] M. Yamaguti and T. Nishida, *On some global solution for quasi-linear hyperbolic equations*, Funkcial. Ekvac., 11 (1968), pp. 51–57.

SIAM J. MATH. ANAL.
Vol. 32, No. 2, pp. 291–303

ON THE SPATIAL DECAY OF SOLUTIONS TO A QUASI-LINEAR PARABOLIC INITIAL-BOUNDARY VALUE PROBLEM AND THEIR DERIVATIVES*

L. E. PAYNE† AND G. A. PHILIPPIN‡

Abstract. In this paper, we study the spatial decay of the solution of a quasi-linear heat equation in a long cylindrical region if the far end and the lateral surface are held at a zero temperature and a nonzero temperature is applied at the near end. Our result follows from the maximum principle applied to an auxiliary function Φ defined on the solution u and its derivatives.

Key words. parabolic equations, decay bounds, maximum principle

AMS subject classifications. 35K55, 35K60, 35B50

PII. S0036141099356519

1. Introduction. In this paper, we are interested in the spatial decay of the solution of a quasi-linear heat equation in a long cylindrical region if the far end and the lateral surface are held at zero temperature and a nonzero temperature is applied at the near end. The specific equation we consider is of the form

$$\Delta u - \epsilon(x_3, t)\,\frac{\partial u}{\partial t} + f(u) = 0$$

in $\Omega \times \mathbb{R}^+$, where the generators of the cylinder Ω are parallel to the x_3-axis, Δ denotes the Laplace operator, and the function $f(u)$ may be such that under certain conditions, the solution can blow up at some point in space time. Our goal is to determine data hypotheses which will ensure that the solution remains bounded and to demonstrate that under these hypotheses, the solution decays at least exponentially in x_3, and when ϵ is constant the solution decays for fixed t at least as fast as $e^{-cx_3^2}$, where c depends on t. In fact, we derive an explicit decay bound for the solution and its cross sectional derivatives.

Decay in time for a related problem has previously been studied by the authors [8]. Some of the arguments here are patterned after arguments in that paper.

In section 2, we establish a maximum principle for a combination of u and its derivatives, and in section 3, we use this principle to derive the explicit decay bounds for u and its cross sectional derivatives. Finally, in section 4, we use the results of section 3 to establish further norm bounds.

2. A maximum principle. Let D be a bounded convex domain in the (x_1, x_2)-plane and let $u(\mathbf{x}, t)$ be defined in the finite cylinder $\Omega := D \times [0, L]$ for $0 < t < T$ as the classical solution of the following initial-boundary value problem:

$$\Delta u - \epsilon(x_3, t)\,\frac{\partial u}{\partial t} + f(u) = 0, \quad \mathbf{x} := (x_1, x_2, x_3) \in \Omega, \quad 0 < t < T, \tag{2.1}$$

$$u(\mathbf{x}, t) = 0, \quad \mathbf{x} \in \partial\Omega_L \cup \partial\Omega_{\text{lat}}, \quad 0 < t < T, \tag{2.2}$$

*Received by the editors May 24, 1999; accepted for publication October 20, 1999; published electronically June 27, 2000.
http://www.siam.org/journals/sima/32-2/35651.html
†Department of Mathematics, Cornell University, Ithaca, NY 14853 (lep8@cornell.edu).
‡Département de Mathématiques, Université Laval, Laval, Québec, Canada G1K 7P4 (gphilip@mat.ulaval.ca).

$$u(\mathbf{x},t) = h(x_1,x_2,t), \quad \mathbf{x} \in \partial\Omega_0, \quad 0 < t < T, \tag{2.3}$$

$$u(\mathbf{x},0) = 0, \quad \mathbf{x} \in \Omega, \tag{2.4}$$

with $\partial\Omega_0 := D \times \{0\}, \partial\Omega_L := D \times \{L\}, \partial\Omega_{\text{lat}} := \partial D \times (0,L)$, $\epsilon(x_3,t) > 0, f'(u) \geq 0, uf(u) \geq 0, f(0) = 0$, and $h(x_1,x_2,t)$ is a prescribed bounded function of its arguments such that $h(x_1,x_2,0) = 0$.

In this section, we derive a maximum principle for the quantity

$$\Phi(\mathbf{x},t) := \left\{ \left(\frac{\partial u}{\partial x_1}\right)^2 + \left(\frac{\partial u}{\partial x_2}\right)^2 + au^2 + 2F(u) \right\} g(x_3,t), \tag{2.5}$$

defined on the solution of (2.1), (2.2), (2.3), (2.4), where $F(u)$ and $g(x_3,t)$ are to be appropriately chosen functions and a is a nonnegative constant.

It is well known that the solution of (2.1), (2.2), (2.3), (2.4) will exist up to time of blow-up [1], [5], [6] if in fact the solution does blow up. We shall first assume that T is some time prior to the blow-up time. In section 3, we will demonstrate that if the data are appropriately restricted, the solution will remain bounded for all time and that in this case T may be taken to be infinity. We first establish the following lemma.

LEMMA 2.1. *Let u be a classical solution of* (2.1) *with $f' \geq 0$. Let $g(x_3,t)$ be a particular positive solution of the parabolic inequality*

$$\epsilon(x_3,t)\left(\frac{1}{g}\right)_{,t} - \left(\frac{1}{g}\right)_{,x_3x_3} + \frac{2a}{g} \geq 0. \tag{2.6}$$

Let $F(u)$ be defined as

$$F(u) := \int_0^u f(\eta)\, d\eta. \tag{2.7}$$

Assume moreover that

$$uF' - 2F \geq 0. \tag{2.8}$$

We then conclude that $\Phi(\mathbf{x},t)$ defined in (2.5) *satisfies the following parabolic inequality:*

$$L\Phi := \Delta\Phi - \epsilon(x_3,t)\,\Phi_{,t} + (u_{,\alpha}u_{,\alpha})^{-1} w_\beta \Phi_{,\beta} \geq 0, \tag{2.9}$$

$\mathbf{x} \in \Omega$, $0 < t < T$, where w_β are functions regular throughout $\Omega \times (0,T)$.

In Lemma 2.1 and in the remainder of the paper, we adopt the following notation: $u_{,\alpha} := \frac{\partial u}{\partial x_\alpha}$, $\alpha = 1,2$; $u_{,k} := \frac{\partial u}{\partial x_k}$, $k = 1,2,3$; $u_{,t} := \frac{\partial u}{\partial t}$. Moreover, summation over repeated indices is assumed from 1 to 2 for Greek indices and from 1 to 3 for Latin indices.

For the proof of Lemma 2.1, we compute

$$\Phi_{,\beta} = (2u_{,\alpha\beta}u_{,\alpha} + 2auu_{,\beta} + 2F'u_{,\beta})g, \tag{2.10}$$

$$\begin{aligned}\Phi_{,\beta\beta} = {} & (2u_{,\alpha\beta\beta}u_{,\alpha} + 2u_{,\alpha\beta}u_{,\alpha\beta} + 2au_{,\beta}u_{,\beta} + 2auu_{,\beta\beta} \\ & + 2F''u_{,\beta}u_{,\beta} + 2F'u_{,\beta\beta})\, g,\end{aligned} \tag{2.11}$$

$$\Phi_{,x_3} = (2u_{,\alpha}u_{,\alpha x_3} + 2auu_{,x_3} + 2F'u_{,x_3})\, g + \Phi\frac{g_{,x_3}}{g}, \tag{2.12}$$

$$\begin{aligned}\Phi_{,x_3x_3} = {} & (2u_{,\alpha x_3}u_{,\alpha x_3} + 2u_{,\alpha}u_{,\alpha x_3x_3} + 2au_{,x_3}^2 + 2auu_{,x_3x_3} \\ & +2F''u_{,x_3}^2 + 2F'u_{,x_3x_3})\, g \\ & + \left(\Phi_{,x_3} - \Phi\,\frac{g_{,x_3}}{g}\right)\frac{g_{,x_3}}{g} + \Phi_{,x_3}\,\frac{g_{,x_3}}{g} + \Phi\left(\frac{g_{,x_3}}{g}\right)_{,x_3},\end{aligned} \tag{2.13}$$

$$\begin{aligned}\Delta\Phi = {} & \{2u_{,\alpha kk}u_{,\alpha} + 2u_{,\alpha k}u_{,\alpha k} + 2a|\nabla u|^2 + 2au\Delta u \\ & +2F''|\nabla u|^2 + 2F'\Delta u\}\, g \\ & +2\Phi_{,x_3}\,\frac{g_{,x_3}}{g} + \Phi\left\{\left(\frac{g_{,x_3}}{g}\right)_{,x_3} - \left(\frac{g_{,x_3}}{g}\right)^2\right\},\end{aligned} \tag{2.14}$$

$$\Phi_{,t} = (2u_{,\alpha t}u_{,\alpha} + 2auu_{,t} + 2F'u_{,t})\, g + \Phi\,\frac{g_{,t}}{g}. \tag{2.15}$$

Combining (2.14) and (2.15) leads to

$$\begin{aligned}& \Delta\Phi - \epsilon(x_3,t)\,\Phi_{,t} \\ & = \{2u_{,\alpha k}u_{,\alpha k} + 2a|\nabla u|^2 + 2u_{,\alpha}[\Delta u - \epsilon u_{,t}]_{,\alpha} \\ & \quad +2(F' + au)[\Delta u - \epsilon u_{,t}] + 2F''|\nabla u|^2\}\, g \\ & \quad +2\Phi_{,x_3}\,\frac{g_{,x_3}}{g} + \Phi\left\{\left(\frac{g_{,x_3}}{g}\right)_{,x_3} - \left(\frac{g_{,x_3}}{g}\right)^2 - \epsilon\,\frac{g_{,t}}{g}\right\}.\end{aligned} \tag{2.16}$$

The first two terms on the right-hand side of (2.16) may be estimated as follows:

$$\begin{aligned}2u_{,\alpha k}u_{,\alpha k} + 2a|\nabla u|^2 & \geq 2u_{,\alpha\beta}u_{,\alpha\beta} + 2au_{,\alpha}u_{,\alpha} \\ & \geq \frac{2u_{,\alpha\beta}u_{,\beta}u_{,\alpha\gamma}u_{,\gamma}}{u_{,\delta}u_{,\delta}} + 2au_{,\alpha}u_{,\alpha},\end{aligned} \tag{2.17}$$

where we have used the Schwarz inequality at the last step of (2.17). Making iterated use of (2.10), we may represent $u_{,\alpha\beta}u_{,\beta}u_{,\alpha\gamma}u_{,\gamma}$ as follows:

$$u_{,\alpha\beta}u_{,\beta}u_{,\alpha\gamma}u_{,\gamma} = 2[au + F']^2u_{,\beta}u_{,\beta} + \widetilde{w}_\beta\Phi_{,\beta}, \tag{2.18}$$

where $\widetilde{w}_\beta$ are two functions regular throughout $\Omega\times(0,\infty)$. Combining (2.16), (2.17), and (2.18) leads to

$$\begin{aligned}& \Delta\Phi - \epsilon(x_3,t)\,\Phi_{,t} + \frac{1}{u_{,\alpha}u_{,\alpha}}\,\Phi_{,\beta}w_\beta \\ & \geq \big\{2[au + F']^2 + 2au_{,\alpha}u_{,\alpha} + 2u_{,\alpha}[\Delta u - \epsilon u_{,t}]_{,\alpha} + 2F''|\nabla u|^2 \\ & \quad +2[F' + au][\Delta u - \epsilon u_{,t}]\big\}\, g + \Phi\left\{\left(\frac{g_{,x_3}}{g}\right)_{,x_3} - \left(\frac{g_{,x_3}}{g}\right)^2 - \epsilon\,\frac{g_{,t}}{g}\right\} \\ & = \{2(F' + au)[\Delta u - \epsilon u_{,t} + F'] + 2u_{,\alpha}[\Delta u - \epsilon u_{,t} + F']_{,\alpha} \\ & \quad +2F''u_{,x_3}^2 + 2a[uF' - 2F]\}\, g + \Phi g\left\{\epsilon\left(\frac{1}{g}\right)_{,t} - \left(\frac{1}{g}\right)_{,x_3x_3} + \frac{2a}{g}\right\},\end{aligned} \tag{2.19}$$

from which the conclusion of Lemma 2.1 follows.

From Lemma 2.1 and Nirenberg's maximum principle [7], [10] it follows then that $\Phi(\mathbf{x},t)$ ($\not\equiv$ const.) takes its maximum value either at $(\mathbf{x},t)\in\partial\Omega\times(0,T]$ or initially at $\mathbf{x}\in\Omega$, or at $(\mathbf{x}^*,t^*)\in\Omega\times\{0<t\le T\}$, at which $u_{,\alpha}u_{,\alpha}=0$. Moreover it follows from Friedman's boundary lemma [2], [10] that Φ ($\not\equiv$ const.) cannot take its maximum value on $\partial\Omega_L\cup\partial\Omega_{\rm lat}$ since the normal derivative of Φ is nonpositive on $\partial\Omega_L\cup\partial\Omega_{\rm lat}$. Indeed, we have

$$\frac{\partial\Phi}{\partial n}=\frac{\partial\Phi}{\partial x_3}=0\quad\text{ on }\partial\Omega_L \tag{2.20}$$

and

$$\frac{\partial\Phi}{\partial n}=2u_nu_{nn}g=-2u_n^2gk\le 0\quad\text{ on }\partial\Omega_{\rm lat}, \tag{2.21}$$

where k is the curvature of ∂D. Finally, it is clear that Φ cannot take its maximum value initially since $\Phi=0$ at time $t=0$.

3. Decay bounds for $\mathbf{u^2}$ and $\mathbf{u_{,\alpha}u_{,\alpha}}$. In this section, we want to eliminate the possibility that Φ defined in (2.5) takes its maximum value at $(\mathbf{x}^*,t^*)\in\Omega\times(0,T]$, where $u_{,\alpha}u_{,\alpha}=0$. This will be achieved by making the parameter a small enough. Under such circumstances, Φ will achieve its maximum value on $\partial\Omega_0\times(0,T]$. Decay bounds for u^2 and $u_{,\alpha}u_{,\alpha}$ are then obtained from the above information.

We first consider the linear case corresponding to $f(u)=0$ in (2.1). In this case, of course, we may take $T=\infty$. Let us assume that Φ takes its maximum value at $(\mathbf{x},t^*)\in\Omega\times\mathbb{R}^+$ at which $u_{,\alpha}u_{,\alpha}=0$, i.e., assume

$$\Phi=\{u_{,\alpha}u_{,\alpha}+au^2\}\,g(x_3,t)\le au_M^2g(x_3^*,t^*), \tag{3.1}$$

where $u_M^2:=\max_{\Omega\times\mathbb{R}^+}(u^2)<\infty$ since the boundary data h are bounded. Evaluated at $x_3=x_3^*$, $t=t^*$, we obtain the inequality

$$u_{,\alpha}u_{,\alpha}(x_1,x_2,x_3^*,t^*)\le a[u_M^2-u^2(x_1,x_2,x_3^*,t^*)]. \tag{3.2}$$

Let $d\ell$ denote the element of length along the straight line in D from (x_1^*,x_2^*) to the nearest point $(\bar x_1,\bar x_2)$ of ∂D. We may then write

$$\left|\frac{du}{d\ell}\right|\le[u_{,\alpha}u_{,\alpha}]^{1/2}\le a^{1/2}[u_M^2-u^2(x_1,x_2,x_3^*,t^*)]^{1/2}. \tag{3.3}$$

Integrating along this line, we obtain

$$\frac{\pi}{2}\le\sqrt{a}|\mathbf{x}^*-\overline{\mathbf{x}}|\le\sqrt{a}d, \tag{3.4}$$

where d is the inradius of D. It then follows from (3.4) that the inequality

$$a\ge\frac{\pi^2}{4d^2}=:a_0 \tag{3.5}$$

is a necessary condition for the maximum of Φ to occur at $(\mathbf{x}^*,t^*)$. If (3.5) is violated, then Φ cannot have its maximum at $(\mathbf{x}^*,t^*)$, and it must therefore occur on $\partial\Omega_0=D\times\{0\}$. We are then led to the inequality

$$\Phi\le\max_{\partial\Omega_0\times\mathbb{R}^+}\Phi, \tag{3.6}$$

valid for all $a \in [0, a_0)$. Increasing a to a_0, we are led to the following result.

THEOREM 3.1. *Let u be the classical solution of* (2.1), (2.2), (2.3), (2.4) *with $T = \infty$ and $f \equiv 0$. Let $g(x_3, t) > 0$ be any particular solution of* (2.6). *Then the following inequality holds:*

$$u_{,\alpha}u_{,\alpha} + a_0 u^2 \leq \frac{H^2}{g(x_3, t)} \tag{3.7}$$

with

$$H^2 := \max_{D\times\mathbb{R}^+} \{h_{,\alpha}h_{,\alpha} + a_0 h^2]\, g(0, t)\}, \tag{3.8}$$

where a_0 is given by (3.5).

One possible choice of $g(x_3, t)$ is

$$g^{(1)}(x_3) = e^{\sqrt{2a}x_3}. \tag{3.9}$$

However, in the special case $\epsilon(x_3, t) = 1$, there are other simple choices for g. Note that the case $\epsilon = \text{const.}$ reduces to the case $\epsilon = 1$ by rescaling in t. Thus, in the remainder of the paper, we consider the particular case $\epsilon(x_3, t) = 1$ and observe that we may make the following choices for g:

$$g^{(2)}(t) = e^{2at}, \tag{3.10}$$

$$g^{(3)}(x_3, t) = e^{2at}(t + t_0)^{1/2} \exp\{x_3^2/4(t + t_0)\}, \tag{3.11}$$

where t_0 is an arbitrary positive constant in (3.11). With the choice (3.11) for g inequality (3.7) clearly implies exponential decay in t with fixed x_3 for both u^2 and $u_{,\alpha}u_{,\alpha}$. It also implies that for fixed t, u^2 and $u_{,\alpha}u_{,\alpha}$ both decay $0(e^{-cx_3^2})$ for some $c(t)$ as x_3 increases.

Furthermore, the pointwise decay bound is explicit. It was already shown (see [9]) that this rate of spatial decay holds for u itself, but the methods used in [9] do not carry over to the nonlinear problem (2.1), (2.2), (2.3), (2.4). The decay estimates require of course that the data be such that the quantity H^2 in (3.8) is bounded. In the linear case with $T = \infty$, this implies appropriate decay of h^2 and $h_{,\alpha}h_{,\alpha}$ as $t \longrightarrow \infty$.

It should be remarked that in the case $f(u) \equiv 0$ but $\epsilon(x_3, t)$ is an arbitrary nonnegative function of x_3 and t, the choice (3.9) would imply exponential decay in x_3 for both u^2 and $u_{,\alpha}u_{,\alpha}$, in which case we would have

$$H^2 := \max_{D\times\mathbb{R}^+} \{h_{,\alpha}h_{,\alpha} + a_0 h^2\}.$$

We turn now to the nonlinear case, i.e., the case of (2.1) with $f \not\equiv 0$. For simplicity, we assume that the boundary data $h(x_1, x_2, t)$ are nonnegative so that u will also be nonnegative. Moreover, we assume that $f(s)$ is differentiable on $(0, \infty)$ and satisfies the conditions

$$sf'(s) \geq f(s) > 0, \quad s > 0, \quad f(0) = 0. \tag{3.12}$$

Clearly (3.12) implies (2.8) with $f = F'$ and implies that the positive quantity $\frac{f(s)}{s}$ is nondecreasing in s.

As indicated earlier, the solution of (2.1), (2.2), (2.3), (2.4) with $f \geq 0$ may blow up at some time $\hat{t}$ which may be finite or infinite [1], [5], [6]. However, if blow-up does occur at $\hat{t}$, then the solution of (2.1), (2.2), (2.3), (2.4) will exist in $(0, \hat{t})$ and Lemma 2.1 holds true in $(0, T)$ for $T < \hat{t}$, so that Φ defined in (2.5) with $f = F'$ will assume its maximum value either on $\partial\Omega_0 := D \times \{0\}$ or at an interior point $\mathbf{x}^*$ of Ω at $t^* < T$ where $u_{,\alpha}u_{,\alpha} = 0$. Let us assume this latter situation, i.e., assume the inequality

$$\Phi(\mathbf{x}, t) \leq \Phi(\mathbf{x}^*, t^*). \tag{3.13}$$

Evaluating (3.13) at $x_3 = x_3^*$, $t = t^*$, we obtain the inequality

$$u_{,\alpha}u_{,\alpha} \leq a[u^2(\mathbf{x}^*, t^*) - u^2] + 2[F(u(\mathbf{x}^*, t^*)) - F(u)]. \tag{3.14}$$

Using the generalized mean value theorem and (3.12), we may write

$$\begin{aligned} 2[F(u(\mathbf{x}^*, t^*)) - F(u)] &= \frac{f(s)}{s}[u^2(\mathbf{x}^*, t^*) - u^2] \\ &\leq \frac{f(u(\mathbf{x}^*, t^*))}{u(\mathbf{x}^*, t^*)}[u^2(\mathbf{x}^*, t^*) - u^2], \end{aligned} \tag{3.15}$$

where s is some intermediate value in $(u, u(\mathbf{x}^*, t^*))$. Combining (3.14) and (3.15), we obtain

$$u_{,\alpha}u_{,\alpha} \leq \left\{ a + \frac{f(u(\mathbf{x}^*, t^*))}{u(\mathbf{x}^*, t^*)} \right\} [u^2(\mathbf{x}^*, t^*) - u^2], \tag{3.16}$$

from which we obtain the inequality

$$a \geq \frac{\pi^2}{4d^2} - \frac{f(u(\mathbf{x}^*, t^*))}{u(\mathbf{x}^*, t^*)}, \tag{3.17}$$

in analogy to (3.5). As before, we then conclude that any nonzero Φ cannot reach its maximum at $(\mathbf{x}^*, t^*)$ if we select

$$0 \leq a < a_1 := \frac{\pi^2}{4d^2} - \frac{f(u(\mathbf{x}^*, t^*))}{u(\mathbf{x}^*, t^*)}. \tag{3.18}$$

The value a_1 defined by (3.18) is, however, of no practical use. We obviously need some conditions on the boundary data h which lead to an explicitly computable upper bound for $\frac{f(u)}{u}$ which is strictly dominated by $\frac{\pi^2}{4d^2}$. Such conditions are given in the next lemma.

LEMMA 3.2. *Let $\phi(x_1, x_2)$ be the first Dirichlet eigenfunction of D, i.e.,*

$$\phi_{,\alpha\alpha} + \lambda\phi = 0, \quad \phi > 0 \text{ in } D, \quad \phi = 0 \text{ on } \partial D, \tag{3.19}$$

normalized by

$$\max_D \phi = 1. \tag{3.20}$$

Let x_0, a, M be positive constants such that

$$0 \leq h(x_1, x_2, t) \leq M\phi \frac{1}{\sqrt{t}} \exp\left(-\frac{x_0^2}{4t} - at \right). \tag{3.21}$$

Let $\hat{h}$ *be defined as*

$$\hat{h} := M \max_{t>0} \left\{ \frac{1}{\sqrt{t}} \exp\left(-\frac{x_0^2}{4t} - at \right) \right\} \tag{3.22}$$
$$= \frac{M}{x_0} \left\{ 1 + [4ax_0^2 + 1]^{1/2} \right\}^{1/2} \exp\left(-\frac{1}{2}(4ax_0^2 + 1)^{1/2} \right).$$

Assume (3.12) *and that* $\hat{h}$ *is such that*

$$\frac{f(\hat{h})}{\hat{h}} < \frac{\pi^2}{4d^2} - a. \tag{3.23}$$

Then we conclude that $\frac{f(u)}{u}$ *remains bounded away from* $\frac{\pi^2}{4d^2} - a$ *for all time, i.e., we have*

$$\frac{f(u(\mathbf{x},t))}{u(\mathbf{x},t)} < \frac{\pi^2}{4d^2} - a, \quad \mathbf{x} \in \Omega, \quad t > 0. \tag{3.24}$$

Moreover, we have the following estimate:

$$u(\mathbf{x},t) \le U(\mathbf{x},t) := \frac{M\phi}{\sqrt{t}} \exp\left(-\frac{(x_0 + x_3)^2}{4t} - at \right), \mathbf{x} \in \Omega, t > 0. \tag{3.25}$$

For the proof of Lemma 3.2, we let U satisfy

$$\Delta U - \frac{\partial U}{\partial t} + (\lambda - a)\, U = 0 \quad \text{in } \Omega \times \mathbb{R}^+, \tag{3.26}$$

$$U \ge h \quad \text{on } \partial\Omega_0 \times \mathbb{R}^+, \tag{3.27}$$

$$U \ge 0 \quad \text{on } \partial\Omega_L \times \mathbb{R}^+, \tag{3.28}$$

$$U = 0 \quad \text{on } \partial\Omega_{\text{lat}} \times \mathbb{R}^+, \tag{3.29}$$

$$U(\mathbf{x},t) = 0 \quad \text{in } \Omega, \quad t \longrightarrow 0. \tag{3.30}$$

Suppose now that (3.24) is violated and let $\tilde{t}$ be the first time for which $\frac{f(u)}{u}$ reaches the value $\frac{\pi^2}{4d^2} - a$ at some point $\tilde{\mathbf{x}} \in \Omega$. Then we would have

$$\max_{\mathbf{x}\in\Omega} \frac{f(u(\mathbf{x},t))}{u(\mathbf{x},t)} < \frac{\pi^2}{4d^2} - a \quad \text{for all } t \in [0,\tilde{t}). \tag{3.31}$$

From (3.26), (3.31), and the inequality

$$\lambda > \frac{\pi^2}{4d^2} \tag{3.32}$$

established by Hersch in [3], we then obtain

$$\Delta U - \frac{\partial U}{\partial t} < -U \max_{\mathbf{x}\in\Omega} \frac{f(u(\mathbf{x},t))}{u(\mathbf{x},t)}, \quad t \in [0,\tilde{t}). \tag{3.33}$$

Setting

$$w := U - u, \tag{3.34}$$

we obtain from (2.1) and (3.33)

$$\Delta w - \frac{\partial w}{\partial t} < -w \max_{\Omega} \frac{f(u(\mathbf{x},t))}{u(\mathbf{x},t)}, \quad t \in [0,\tilde{t}). \tag{3.35}$$

It then follows from the maximum principle that $w(\mathbf{x},t)$ is positive in $\Omega \times (0,\tilde{t})$, i.e., we have (3.25) in $(0,\tilde{t})$, and we obtain using (3.20), (3.23)

$$u(\mathbf{x},t) \le \hat{h}, \quad t \in (0,\tilde{t}). \tag{3.36}$$

Finally, (3.36), (3.12), and (3.23) imply the inequality

$$\frac{f(u)}{u} \le \frac{f(\hat{h})}{\hat{h}} < \frac{\pi^2}{4d^2} - a, \quad 0 \le t \le \hat{t}. \tag{3.37}$$

In particular, we have

$$\frac{f(u(\mathbf{x},\tilde{t}))}{u(\tilde{\mathbf{x}},\tilde{t})} < \frac{\pi^2}{4d^2} - a, \tag{3.38}$$

in contradiction to the definition of $(\tilde{\mathbf{x}},\tilde{t})$. We then conclude that $\tilde{t} = \infty$, and the proof of Lemma 3.2 is complete.

As a consequence of Lemma 3.2 together with the assertion that includes (3.18), we obtain exponential decay bounds for u^2 and for $u_{,\alpha}u_{,\alpha}$ given in the next theorem.

THEOREM 3.3. *Under the assumptions of Lemma* 3.2*, let us define*

$$a_1 := \frac{\pi^2}{4d^2} - \frac{f(\hat{h})}{\hat{h}}. \tag{3.39}$$

We then conclude that Φ *given by* (2.5), (3.11) *takes its maximum value on* $\partial\Omega_0 \times (0,t)$ *for all* $a \in [0,a_1)$*. With* $a \longrightarrow a_1$*, we obtain*

$$u_{,\alpha}u_{,\alpha} + a_1 u^2 + 2F(u) \le Q^2(x_3,t) \tag{3.40}$$

with

$$Q^2(x_3,t) := \frac{\mathcal{H}^2}{\sqrt{t+t_0}} \exp\left(-2a_1 t - \frac{x_3^2}{4(t+t_0)}\right) \tag{3.41}$$

and

$$\mathcal{H}^2 := \max_{D\times\mathbb{R}^+} \left\{ [h_{,\alpha}h_{,\alpha} + a_1 h^2 + 2F(h)]\, e^{2a_1 t}\sqrt{t+t_0} \right\}. \tag{3.42}$$

We note that for $t \in \mathbb{R}^+$, (3.21) implies that h must decay at least exponentially in t. However, if we are interested only in a finite time interval $(0,\bar{t})$, then we may take

$$M := \max_{D\times(0,\bar{t})} \left\{ h(x_1,x_2,t)\, t^{1/2} \exp\left(at + \frac{x_0^2}{4t}\right) \right\}. \tag{3.43}$$

Thus with the quantity $\hat{h}$ of (3.22) defined using this value of M and (3.23) satisfied for this modified $\hat{h}$, (3.40) holds for $\mathcal{H}^2$ defined by (3.42) except that now the maximum is taken over $D \times (0,\bar{t}\,)$. It follows from (3.40) that

$$u(\mathbf{x},t) \le \frac{Q(x_3,t)}{\sqrt{a_1}}. \tag{3.44}$$

However, a sharper estimate for $u(\mathbf{x},t)$ may be derived as follows: From (3.40), we have

$$\left|\frac{\partial u}{\partial \ell}\right| \le (u_\alpha u_\alpha)^{1/2} \le \{Q^2(x_3,t) - a_1 u^2\}^{1/2}, \tag{3.45}$$

where $d\ell$ is the element of length along the line joining $\mathbf{P} := (x_1,x_2,x_3,t)$ in $\Omega \times \mathbb{R}^+$ to the nearest point $\tilde{\mathbf{P}} := (\tilde{x}_1,\tilde{x}_2,x_3,t)$ on $\partial\Omega_{\text{lat}} \times \mathbb{R}^+$. Integrating along this line, we obtain

$$\int_0^u \frac{d\eta}{\sqrt{Q^2 - a_1\eta^2}} \le \delta, \tag{3.46}$$

where δ is the distance of $\mathbf{P}$ from the lateral surface of the cylinder. It follows from (3.46) that

$$u \le \frac{Q(x_3,t)}{\sqrt{a_1}} \sin(\sqrt{a_1}\delta) \le \delta Q(x_3,t). \tag{3.47}$$

This estimate reflects the fact that u is small near $\partial\Omega_{\text{lat}}$.

4. Further norm bounds. Clearly a bound for the quantity

$$\Psi(x_3,t) := \sqrt{\int_D u^2 dx_1 dx_2} \tag{4.1}$$

may be obtained by integration of (3.47). However, a bound which imposes somewhat less restrictive hypotheses on the data may be derived from the maximum principle together with the following lemma.

LEMMA 4.1. *Under the assumptions of Lemma* 3.2*, we have the following parabolic inequality for the quantity* $\Psi(x_3,t)$*:*

$$\Psi_{,x_3x_3} - \Psi_{,t} \ge a\Psi, \quad x_3 \in [0,L], \quad t > 0, \tag{4.2}$$

with

$$a := \lambda - \max_{\Omega\times\mathbb{R}^+} \frac{f(u)}{u}, \tag{4.3}$$

where λ *is the first eigenvalue of* D*.*

It follows from (3.32) and (3.37) that a in (4.3) is positive. For the proof of (4.2), we compute

$$\Psi_{,x_3} = \frac{1}{\Psi}\int_D u u_{,x_3} dx_1 dx_2, \tag{4.4}$$

$$\begin{aligned}\Psi_{,x_3x_3} &= \frac{1}{\Psi}\int_D \left(u_{,x_3}^2 + u u_{,x_3x_3}\right) dx_1 dx_2 - \frac{1}{\Psi^3}\left\{\int_D u u_{,x_3} dx_1 dx_2\right\}^2 \\ &\ge \frac{1}{\Psi}\int_D u u_{,x_3x_3} dx_1 dx_2,\end{aligned} \tag{4.5}$$

$$\Psi_{,t} = \frac{1}{\Psi}\int_D u u_{,t} dx_1 dx_2. \tag{4.6}$$

Combining (4.5) and (4.6), we obtain

$$\Psi_{,x_3x_3} - \Psi_{,t} \geq \frac{1}{\Psi}\int_D u\left[u_{,x_3x_3} - u_{,t}\right]\, dx_1dx_2 \tag{4.7}$$
$$= -\frac{1}{\Psi}\int_D u\left[u_{,\alpha\alpha} + f(u)\right]\, dx_1dx_2$$
$$\geq \frac{1}{\Psi}\int_D u^2\left[\lambda - \frac{f(u)}{u}\right]\, dx_1dx_2 \geq a\Psi,$$

which is the desired inequality.

From (4.2) and the maximum principle, we obtain

$$\Psi(x_3,t) \leq \hat{M}(t+t_0)^{-1/2}\exp\left\{-at - x_3^2/4(t+t_0)\right\} \tag{4.8}$$

with an arbitrary constant $t_0 > 0$ and with

$$\hat{M} := \sup_{\mathbb{R}^+}\left\{e^{at}(t+t_0)^{1/2}\left(\int_D h^2 dx_1dx_2\right)^{1/2}\right\}. \tag{4.9}$$

As before, we note that on any finite time interval $[0,\bar{t}\,]$, we may replace (4.3) by

$$a := \lambda - \max_{\Omega\times[0,\bar{t}\,]}\frac{f(u)}{u} \tag{4.10}$$

in Lemma 4.1, and take the supremum over $[0,\bar{t}\,]$ in (4.9).

We note that a bound for Ψ would also follow from Lemma 4.1 and the arguments of Horgan, Payne, and Wheeler [4].

Our explicit decay bounds have not included bounds for the x_3-derivative of u. However, it is possible to derive an explicit decay bound for the Dirichlet integral $D_z(u)$ defined as

$$D_z(u) := \int_0^t\int_z^L\int_D |\nabla u|^2 d\mathbf{x}d\tau. \tag{4.11}$$

We first establish the following result.

LEMMA 4.2. *Under the assumptions of Lemma* 3.2*, the quantity* $D_z(u)$ *defined in* (4.11) *satisfies the differential inequality*

$$D_z(u) \leq \frac{1}{2a\sqrt{\lambda}}\frac{d}{dz}D_z(u), \tag{4.12}$$

where a *and* λ *have the same meaning as in Lemmas* 3.2 *and* 4.1.

Integrating (4.12)*, we obtain*

$$D_z(u) \leq D_0(u)e^{-2a\sqrt{\lambda}z}. \tag{4.13}$$

For the proof of Lemma 4.2, we write

$$\left(uu_{,\alpha}\right)_{,\alpha} + \left(uu_{,x_3}\right)_{,x_3} = |\nabla u|^2 + u[u_{,t} - f(u)]. \tag{4.14}$$

Integrating (4.14) and using the divergence theorem, we obtain

$$D_z(u) = \int_0^t\int_z^L\int_D uf(u)d\mathbf{x}d\tau - \int_0^t\int_D uu_{,x_3}dx_1dx_2d\tau\Big|_{x_3=z} \tag{4.15}$$
$$-\frac{1}{2}\int_z^L\int_D u^2 dx_1dx_2\Big|_{\tau=t}.$$

In (4.15), the first two terms may be bounded as follows:

$$\int_0^t \int_z^L \int_D uf(u) d\mathbf{x} d\tau \le \left(\max_{\Omega \times \mathbb{R}^+} \frac{f(u)}{u} \right) \int_0^t \int_z^L \int_D u^2 d\mathbf{x} d\tau \le \frac{1}{\lambda} \left(\max_{\Omega \times \mathbb{R}^+} \frac{f(u)}{u} \right) D_z(u), \tag{4.16}$$

$$\begin{aligned} -\int_0^t \int_D u u_{,x_3} dx_1 dx_2 d\tau &\le \left\{ \int_0^t \int_D u^2 dx_1 dx_2 d\tau \int_0^t \int_D u_{,x_3}^2 dx_1 dx_2 d\tau \right\}^{1/2} \\ &\le \frac{1}{2\sqrt{\lambda}} \left\{ \int_0^t \int_D u_{,\alpha} u_{,\alpha} dx_1 dx_2 d\tau + \int_0^t \int_D u_{,x_3}^2 dx_1 dx_2 d\tau \right\} \\ &= \frac{1}{2\sqrt{\lambda}} \int_0^t \int_D |\nabla u|^2 dx_1 dx_2 d\tau = -\frac{1}{2\sqrt{\lambda}} \frac{d}{dz} D_z(u) \end{aligned} \tag{4.17}$$

at $x_3 = z$. We are then led to the differential inequality

$$D_z(u) \left[1 - \frac{1}{\lambda} \left(\max_{\Omega \times \mathbb{R}^+} \frac{f(u)}{u} \right) \right] \le \frac{1}{2\sqrt{\lambda}} \frac{d}{dz} D_z(u), \tag{4.18}$$

which leads to the desired inequality (4.12) since we have (3.24) and (3.32).

As it stands, (4.13) is not explicit since $D(u) := D_0(u)$ is not defined in terms of data alone. However, we may derive a bound for $D(u)$ as follows: Let $v(\mathbf{x}, t)$ be an arbitrary Dirichlet integrable function. Then by the triangle inequality, we have

$$\sqrt{D(u)} \le \sqrt{D(u-v)} + \sqrt{D(v)}. \tag{4.19}$$

Let us choose v to be the harmonic function that for each t satisfies

$$\Delta v = 0 \quad \text{in } \Omega \times \mathbb{R}^+, \tag{4.20}$$

$$v = 0 \quad \text{on } (\partial\Omega_L \cup \partial\Omega_{\text{lat}}) \times \mathbb{R}^+, \tag{4.21}$$

$$v = h \quad \text{on } \partial\Omega_0 \times \mathbb{R}^+. \tag{4.22}$$

Making use of the divergence theorem, the Schwarz inequality, and the Rayleigh principle, we compute

$$\begin{aligned} D(u-v) &= -\int_0^t \int_\Omega (u-v)[u_{,\tau} - f(u)] \, d\mathbf{x} d\tau \\ &= -\frac{1}{2} \int_\Omega (u-v)^2 \, d\mathbf{x} \Big|_{\tau=t} - \int_0^t \int_\Omega (u-v) \, v_{,\tau} d\mathbf{x} d\tau \\ &\quad + \int_0^t \int_\Omega (u-v) \, f(u) \, d\mathbf{x} d\tau \\ &\le \left\{ \int_0^t \int_\Omega (u-v)^2 \, d\mathbf{x} d\tau \right\}^{1/2} \end{aligned}$$

$$\times \left\{ \left[\int_0^t \int_\Omega v_{,\tau}^2 d\mathbf{x} d\tau \right]^{1/2} + \left[\int_0^t \int_D f^2(u)\, d\mathbf{x} d\tau \right]^{1/2} \right\}$$

$$\leq \frac{1}{\sqrt{\lambda}} \sqrt{D(u-v)}$$

$$\times \left\{ \left[\int_0^t \int_\Omega v_{,\tau}^2\, d\mathbf{x} d\tau \right]^{1/2} + \left(\max_{\Omega \times \mathbb{R}^+} \left(\frac{f(u)}{u} \right) \right) \left[\int_0^t \int_\Omega u^2\, d\mathbf{x} d\tau \right]^{1/2} \right\}. \tag{4.23}$$

Moreover, using the triangle inequality and the Rayleigh principle, we may write

$$\sqrt{\int_0^t \int_\Omega u^2\, d\mathbf{x} d\tau} \leq \sqrt{\int_0^t \int_\Omega (u-v)^2\, d\mathbf{x} d\tau} + \sqrt{\int_0^t \int_\Omega v^2\, d\mathbf{x} d\tau}$$

$$\leq \frac{1}{\sqrt{\lambda}} \sqrt{D(u-v)} + \sqrt{\int_0^t \int_\Omega v^2\, d\mathbf{x} d\tau}. \tag{4.24}$$

Combining (4.23) and (4.24), and taking (3.23), (3.32) into account, we obtain

$$\frac{a}{\sqrt{\lambda}} \sqrt{D(u-v)} \leq \sqrt{\int_0^t \int_\Omega v_{,\tau}^2\, d\mathbf{x} d\tau} + \frac{f(\hat{h})}{\hat{h}} \sqrt{\int_0^t \int_\Omega v^2\, d\mathbf{x} d\tau}, \tag{4.25}$$

where $\hat{h}$ is defined in (3.22). It is well known that the L_2 integral of a harmonic function is boundable in terms of the L_2 integral of its Dirichlet data (see, e.g., Sigillito [11]). Thus for computable constants A_1 and A_2, we have

$$\sqrt{D(u-v)} \leq A_1 \sqrt{\int_0^t \int_D h_{,\tau}^2\, dx_1 dx_2 d\tau} + A_2 \sqrt{\int_0^t \int_D h^2\, dx_1 dx_2 d\tau}. \tag{4.26}$$

Finally, for computable A_3 and A_4 (see Sigillito [11]), we may write

$$D(v) \leq A_3 \int_0^t \int_D h^2\, dx_1 dx_2 d\tau + A_4 \int_0^t \int_D h_{,\alpha} h_{,\alpha}\, dx_1 dx_2 d\tau. \tag{4.27}$$

Combining (4.19), (4.26), and (4.27), we obtain

$$D(u) \leq \widetilde{D} := B_1 \int_0^t \int_D h^2\, dx_1 dx_2 d\tau + B_2 \int_0^t \int_D h_{,t}^2\, dx_1 dx_2 d\tau$$

$$+ B_3 \int_0^t \int_D h_{,\alpha} h_{,\alpha}\, dx_1 dx_2 d\tau, \tag{4.28}$$

where B_k are computable constants, $k = 1, 2, 3$. Inequalities (4.13) and (4.28) provide the following explicit decay bound for $D_z(u)$:

$$D_z(u) \leq \widetilde{D} e^{-2a\sqrt{\lambda} z}. \tag{4.29}$$

REFERENCES

[1] J. M. Ball, *Remarks on blow-up and nonexistence theorems for nonlinear evolution equations*, Quart. J. Math. Oxford, 28 (1977), pp. 473–486.

[2] A. Friedman, *Remarks on the maximum principle for parabolic equations and its applications*, Pacific J. Math., 8 (1958), pp. 201–211.

[3] J. Hersch, *Sur la fréquence fondamentale d'une membrane vibrante: évaluations par défaut et principe du maximum*, Z. Angew. Math. Phys., 11 (1960), pp. 387–413.

[4] C. O. Horgan, L. E. Payne, and L. T. Wheeler, *Spatial decay estimates in transient heat conduction*, Quart. Appl. Math., 42 (1984), pp. 119–127.

[5] H. Kielhöfer, *Halbgruppen und semilineare Anfangs-Randwert-probleme*, Manuscripta Math., 12 (1974), pp. 121–152.

[6] H. A. Levine, *Some nonexistence and instability theorems for solutions of formally parabolic equations of the form* $Pu_t = -Au + f(u)$, Arch. Rational Mech. Anal., 51 (1973), pp. 371–386.

[7] L. Nirenberg, *A strong maximum principle for parabolic equations*, Comm. Pure Appl. Math., 6 (1953), pp. 167–177.

[8] L. E. Payne and G. A. Philippin, *Decay bounds for solutions of second order parabolic problems and their derivatives*, Math. Models Methods Appl. Sci., 5 (1995), pp. 95–110.

[9] L. E. Payne and G. A. Philippin, *Pointwise bounds and spatial decay estimates in heat conduction problems*, Math. Models Methods Appl. Sci., 5 (1995), pp. 755–775.

[10] M. H. Protter and H. F. Weinberger, *Maximum Principles in Differential Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1967.

[11] V. G. Sigillito, *Explicit A Priori Inequalities with Applications to Boundary Value Problems*, Pitman Res. Notes Math. Ser. 13, Longman, Harlow, UK, 1977.

SIAM J. MATH. ANAL.
Vol. 32, No. 2, pp. 304–322

ON A STOCHASTIC HYPERBOLIC SYSTEM IN LINEAR ELASTICITY*

JONG UHN KIM†

Abstract. In this paper we discuss the Cauchy problem for linear elasticity with a space-time white noise forcing term. We show that the solution can be represented by a formula analogous to the Riesz formula for solutions of a wave equation. The solution is a generalized stochastic process and is obtained as the limit of a sequence of ordinary stochastic processes. Our basic tool is the Hilbert space method combined with geometric properties of solutions inherent with a hyperbolic system.

Key words. generalized stochastic process, Gaussian process, white noise, hyperbolic system

AMS subject classifications. 35L15, 35R60, 60H15

PII. S0036141099350377

0. Introduction. In this paper we present new results on the Cauchy problem for a hyperbolic system with a white noise forcing term associated with linear elasticity:

$$u_{tt}(t,x) = A(t,x,D_x)u(t,x) + F(t,x)\xi(t,x) \quad \text{in } (0,\infty)\times R^n; \tag{0.1}$$

$$u(0,x) = 0, \qquad u_t(0,x) = 0, \qquad \text{in } R^n, \tag{0.2}$$

where $x = (x_1, \dots, x_n) \in R^n$, $u = (u_1, \dots, u_n)$ denotes the displacement from equilibrium. $A(t,x,D_x)$ is a second order matrix differential operator, $F(t,x)$ is a matrix function, and $\xi(t,x)$ stands for a vector-valued space-time white noise. Here we only focus on the nonhomogeneous forcing term, and assume the zero initial conditions. Since the system is linear, nonzero initial conditions can be handled separately. The motion of a one-dimensional elastic medium driven by a random noise is typically described by a one-dimensional wave equation with a space-time white noise. This was discussed by Walsh [15] in the framework of Brownian sheets. A numerical result is shown on the front cover of the monograph [7]. Other versions were investigated by Cabaña [1] and Orsingher [13]. It was also discussed by Da Prato and Zabczyk [3] as an example of abstract evolution equations. A semilinear hyperbolic equation with one-dimensional space variable was discussed by Nualart [12] as a two-parameter stochastic differential equation. Marcus and Mizel [10] studied initial-boundary value problems for a stochastic hyperbolic system in one space dimension. A two-dimensional semilinear wave equation was discussed by Dalang and Frangos [4], and Mueller [11] when the random noise is a white noise in time with smooth spatial covariance.

The purpose of this work is to establish the existence of a solution of (0.1) and (0.2) by a representation formula analogous to the Riesz formula for solutions of a wave equation. The representation formula gives good information on the structure of the solution which is a generalized stochastic process. This motivates our quest for the representation formula. Riesz [14] used analytical theory of integrals of fractional

*Received by the editors January 4, 1999; accepted for publication (in revised form) January 24, 2000; published electronically July 5, 2000.
http://www.siam.org/journals/sima/32-2/35037.html
†Department of Mathematics, Virginia Tech, Blacksburg, VA 24061-0123 (kim@math.vt.edu).

order to establish the formula

$$u(t,x) = \frac{1}{2^n \pi^{(n-1)/2}((n-1)/2)!} \Lambda^{(n-1)/2} \int_{\Xi(t,x)} f(s,y)\, dy\, ds, \tag{0.3}$$

where $n \geq 3$ is an odd integer, $x \in R^n$, $t \geq 0$, $\Xi(t,x)$ is the backward light cone defined by $(t-s)^2 \geq |x-y|^2$, $0 \leq s \leq t$. Here, u is the solution of

$$\Lambda u = f \qquad \text{for } x \in R^n,\ t \geq 0; \tag{0.4}$$

$$u(0,x) = 0,\ u_t(0,x) = 0 \qquad \text{for } x \in R^n, \tag{0.5}$$

where $\Lambda = \partial_{tt} - \Delta$, and Δ is the Laplacian.

Gaveau [6] applied this formula to the Cauchy problem for a three-dimensional wave equation with a space-time white noise. For a one-dimensional wave equation, the solution can be represented by a stochastic integral because the integral kernel is locally L^2, which is not true in the higher-dimensional case. Gaveau overcame this difficulty by means of the above Riesz formula, for the above integral in (0.3) can be well-defined as a stochastic integral when $f\, dy\, ds$ is replaced by $f\, \xi(dy\, ds)$ where f is locally bounded and ξ is the white noise. Thanks to the explicit structure of the formula, he could first construct a continuous martingale (according to a partial ordering defined in terms of the backward light cone) through the Riesz integral, and then, a solution was obtained by applying the wave operator to this continuous stochastic process. Hence the resulting solution is a generalized stochastic process. However, the particular formula used in [6], which easily generates a martingale, is valid only for odd space dimensions. He also showed that the Cauchy value of this generalized stochastic process is well-defined by the formula of integration by parts.

This work is an outgrowth of our effort to obtain a similar representation formula for the system of equations in linear elasticity. We will show that the unique solution u of the above Cauchy problem (0.1)–(0.2) in any space dimension can be obtained as a generalized stochastic process in the following form:

$$u(t,x;\omega) = L^d V(t,x;\omega), \tag{0.6}$$

where $L = \partial_{tt} I - A(t,x,D_x)$, I is the $n \times n$ identity matrix, d is the smallest integer larger than $(n/2) - 1$, and V is a continuous Gaussian process which satisfies the property of a domain of dependence. In particular, when $n \geq 3$ is an odd integer, $d = (n-1)/2$. Since the Riesz formula is not available in our case, we have to employ an entirely different argument. The core task is to obtain an integral kernel. For this, we use the Hilbert space method combined with the property of a domain of dependence. We then approximate the space-time white noise by truncating the chaos expansion. The corresponding approximate solution is an ordinary stochastic process, and the true solution is obtained as the limit. Our results are completely new and our approach is different from those of all the previous works.

1. Preliminaries and statement of the main result. Throughout this paper, we make the following assumptions.

The matrix operator $A(t,x,D_x)$ is given by

$$A_{ij}(t,x,D_x) = \sum_{k,l=1}^{n} C_{ij}^{kl}(t,x) \frac{\partial^2}{\partial x_k \partial x_l} + \sum_{k=1}^{n} D_{ij}^{k}(t,x) \frac{\partial}{\partial x_k}$$

for $i, j = 1, \dots, n$. All coefficients are real valued and

(I) $C_{ij}^{kl}(t,x) \in C^\infty(R^{n+1})$ and all the derivatives of each C_{ij}^{kl} are bounded on R^{n+1};

(II) $C_{ij}^{kl}(t,x) = C_{kj}^{il}(t,x) = C_{ji}^{lk}(t,x)$ for all $(t,x) \in R^{n+1}$ and every i, j, k, and l;

(III) there is a positive constant c_0 such that

$$\sum_{k,l,i,j} C_{ij}^{kl}(t,x)\epsilon_{ki}\epsilon_{lj} \geq c_0 \sum_{k,i} \epsilon_{ki}\epsilon_{ki} \tag{1.1}$$

for all $(t,x) \in R^{n+1}$ and every symmetric tensor ϵ_{ki};

(IV) $D_{ij}^{k}(t,x) \in C^\infty(R^{n+1})$ and all the derivatives of each D_{ij}^{k} are bounded on R^{n+1};

(V) $F(t,x)$ is an $n \times n$ matrix function whose components are all measurable and bounded on each bounded subset of R^{n+1}.

We now list some known facts about the following deterministic Cauchy problem:

$$u_{tt}(t,x) = A(t,x,D_x)u(t,x) + f(t,x) \quad \text{in } (0,\infty) \times R^n; \tag{1.2}$$

$$u(0,x) = u_0(x), \qquad u_t(0,x) = u_1(x) \quad \text{in } R^n. \tag{1.3}$$

Let a real number s and a positive number T be given.

THEOREM 1.1. *For given $f(t,x) \in L^2(0,T;(H^s(R^n))^n)$, $u_0(x) \in (H^{s+1}(R^n))^n$, and $u_1(x) \in (H^s(R^n))^n$, there is a unique solution $u(t,x)$ of (1.2) and (1.3) in $C([0,T];(H^{s+1}(R^n))^n) \cap C^1([0,T];(H^s(R^n))^n)$.*

Here $H^s(R^n)$ denotes the usual Sobolev space. For the technical details of the proof, see [9].

Next we fix any $t_0 \geq 0$ and $x_0 \in R^n$ and define for $0 \leq t \leq t_0$, $\epsilon > 0$ and $\eta > 0$,

$$\Gamma_{(t_0,x_0)}(t;\epsilon,\eta) = \{x \in R^n \mid |x - x_0| < \epsilon + \eta(t_0 - t)\}. \tag{1.4}$$

$\Gamma_{(t_0,x_0)}(t;\epsilon,\eta)$ is an n-dimensional ball for each $0 \leq t \leq t_0$, and $\cup_{0\leq t\leq t_0}\{(t, \overline{\Gamma_{(t_0,x_0)}(t;\epsilon,\eta)})\}$ is a truncated cone in R^{n+1}.

THEOREM 1.2. *Let $s = m$ be a nonnegative integer, and $T > 0$ be given. Choose any $(t_0,x_0) \in [0,T] \times R^n$. Then there is $\eta > 0$ depending on the coefficients of $A(t,x,D_x)$, but independent of u, f, (t_0,x_0), and ϵ such that the above solution satisfies*

$$\begin{aligned} &\|u(t_0,\cdot)\|^2_{(H^{m+1}(\Gamma(t_0;\epsilon,\eta)))^n} + \|u_t(t_0,\cdot)\|^2_{(H^m(\Gamma(t_0;\epsilon,\eta)))^n} \\ &\leq M\left(\|u_0\|^2_{(H^{m+1}(\Gamma(0;\epsilon,\eta)))^n} + \|u_1\|^2_{(H^m(\Gamma(0;\epsilon,\eta)))^n} + \int_0^{t_0} \|f(h,\cdot)\|^2_{(H^m(\Gamma(h;\epsilon,\eta)))^n}\, dh\right), \end{aligned} \tag{1.5}$$

where M is a positive constant independent of u, f, (t_0,x_0), and ϵ. Here the subscript (t_0,x_0) of Γ has been suppressed.

This has been established for a general first order hyperbolic system in [2]. The same argument can be applied to a second order hyperbolic system. See [5] and [9].

COROLLARY 1.3. *Suppose that the support of u_0 and u_1 is disjoint from $\overline{\Gamma_{(t_0,x_0)}(0;\epsilon,\eta)}$, and the support of f is disjoint from $\cup_{0\leq t\leq t_0}\{(t, \overline{\Gamma_{(t_0,x_0)}(t;\epsilon,\eta)})\}$ for some $(t_0,x_0) \in [0,\infty) \times R^n$ in Theorem 1.1. Then the support of u is disjoint from $\cup_{0\leq t\leq t_0}\{(t, \overline{\Gamma_{(t_0,x_0)}(t;\epsilon,\eta)})\}$.*

If $s < 0$, this can be proved by approximating the solution by a sequence of solutions with smooth initial data and f, for which (1.5) can be applied.

Next we set up a base probability space for generalized stochastic processes. Let $\mathcal{S} = (\mathcal{S}(R^{n+1}))^n$ be the space of R^n-valued rapidly decreasing C^∞ functions on R^{n+1}, and $\mathcal{S}' = (\mathcal{S}'(R^{n+1}))^n$ be its dual equipped with the weak-star topology. $\mathcal{B}(\mathcal{S}')$ stands for the set of all Borel subsets of $\mathcal{S}'$. By the Bochner–Minlos theorem, there is a probability measure μ on $\mathcal{B}(\mathcal{S}')$ such that

$$\int_{\mathcal{S}'} \exp(\sqrt{-1}\langle \omega\,,\, \phi\rangle_n)\, d\mu(\omega) = \exp(-\|\phi\|^2_{(L^2(R^{n+1}))^n}/2) \tag{1.6}$$

for all $\phi \in \mathcal{S}$. Here, $\langle \cdot\,,\, \cdot\rangle_n$ denotes the duality pairing between $\mathcal{S}'$ and $\mathcal{S}$, i.e.,

$$\langle \omega\ ,\ \phi\rangle_n = \sum_{i=1}^n \langle \omega_i\ ,\ \phi_i\rangle,$$

for $\omega = (\omega_1, \dots, \omega_n)$, $\phi = (\phi_1, \dots, \phi_n)$, where $\langle \cdot\,,\, \cdot\rangle$ denotes the duality pairing between $\mathcal{S}'(R^{n+1})$ and $\mathcal{S}(R^{n+1})$. By (1.6), it follows that for each $f \in (L^2(R^{n+1}))^n$, $\langle \omega, f\rangle_n$ can be defined to be a random variable on the probability space $(\mathcal{S}', \mathcal{B}(\mathcal{S}'), \mu)$, and, for each positive integer k,

$$E(|\langle \omega, f\rangle_n|^{2k}) = \int_{\mathcal{S}'} |\langle \omega, f\rangle_n|^{2k}\, d\mu = \frac{(2k)!}{k!\, 2^k} \|f\|^{2k}_{(L^2(R^{n+1}))^n}. \tag{1.7}$$

For more details, see [7]. The above white noise $\xi(t, x)$ is a vector-valued generalized stochastic process in the sense that

$$\xi_\phi(\omega) = (\langle \omega_1, \phi_1\rangle, \dots, \langle \omega_n, \phi_n\rangle) \tag{1.8}$$

is a vector-valued random variable on $(\mathcal{S}', \mathcal{B}(\mathcal{S}'), \mu)$ for each $\phi \in (C_0^\infty(R^{n+1}))^n$. In the meantime, $F\xi$ is a vector-valued generalized stochastic process defined by

$$(F\xi)_\phi(\omega) = \left(\sum_{j=1}^n \langle \omega_j, F_{1j}\phi_1\rangle, \dots, \sum_{j=1}^n \langle \omega_j, F_{nj}\phi_n\rangle \right), \tag{1.9}$$

for each $\phi \in (C_0^\infty(R^{n+1}))^n$. Next we define for each $i, j = 1, \dots, n$,

$$B_{ij}(x_0, x_1, \dots, x_n; \omega) = \langle \omega_j(y),\ F_{ij}(y)\chi_{(x_0,\dots,x_n)}(y)\rangle. \tag{1.10}$$

Here, $y \in R^{n+1}$ denotes the variable for the duality action $\langle \cdot\,,\, \cdot\rangle$, and

$$\chi_{(x_0,\dots,x_n)}(y_0, \dots, y_n) = \begin{cases} (-1)^k & \text{if } 0 \le y_i < x_i \text{ or } x_i < y_i < 0 \\ & \quad \text{for each } i = 0, 1, \dots, n, \\ 0 & \text{otherwise,} \end{cases} \tag{1.11}$$

where k is the number of negative x_i's. For each bounded subset K of R^{n+1}, it holds that

$$\int_{R^{n+1}} \left|\chi_{(x_0,\dots,x_n)}(y) - \chi_{(\tilde{x}_0,\dots,\tilde{x}_n)}(y)\right|^2 dy \le C_K \left|(x_0, \dots, x_n) - (\tilde{x}_0, \dots, \tilde{x}_n)\right| \tag{1.12}$$

for all $(x_0,\dots,x_n),\ (\tilde{x}_0,\dots,\tilde{x}_n) \in K,$ for some positive constant C_K. Thus it follows from (1.7) that for each bounded subset K of R^{n+1} and each positive integer m,

$$E\Big(\big|B_{ij}(x_0,\dots,x_n;\omega) - B_{ij}(\tilde{x}_0,\dots,\tilde{x}_n;\omega)\big|^{2m}\Big) \le C_{m,K}\big|(x_0,\dots,x_n) - (\tilde{x}_0,\dots,\tilde{x}_n)\big|^m \tag{1.13}$$

for all $(x_0,\dots,x_n),\ (\tilde{x}_0,\dots,\tilde{x}_n) \in K,$ for some positive constant $C_{m,K}$. By a partition of unity and the Kolmogorov continuity theorem, there is a continuous version of $B_{ij}(x_0,\dots,x_n;\omega)$. See [8] and [12] for the multiparameter version of Kolmogorov's theorem. From now on, we always mean this continuous version.

LEMMA 1.4. *For each $\phi = (\phi_1,\dots,\phi_n)$ with $\phi_i \in C_0^\infty(R^{n+1}),\ i = 1,\dots,n,$ it holds that*

$$\text{the } i\text{th component of } (F\xi)_\phi(\omega) = \sum_{j=1}^n \int_{R^{n+1}} (-1)^{n+1} \frac{\partial^{n+1}\phi_i}{\partial x_0 \cdots \partial x_n} B_{ij}(x_0,\dots,x_n;\omega)\, dx_0 \dots dx_n \tag{1.14}$$

for almost all ω.

Proof. Since $B_{ij}(x_0,\dots,x_n;\omega)$ is a continuous version of the process defined by (1.10), $B_{ij}(x_0,\dots,x_n;\omega)$ is continuous in $(x_0,\dots,x_n),$ for each $\omega,$ and there is a subset $\Omega \in \mathcal{B}(\mathcal{S}')$ with $\mu(\Omega) = 1,$ such that for each $\omega \in \Omega,$

$$B_{ij}(r_0,\dots,r_n;\omega) = \langle \omega_j(y),\ F_{ij}(y)\chi_{(r_0,\dots,r_n)}(y)\rangle \tag{1.15}$$

for all $(r_0,\dots,r_n)$ with $r_i =$ a rational number, $i = 0,\dots,n.$

Choose any $\phi \in (C_0^\infty(R^{n+1}))^n$ and set

$$\psi = \frac{\partial^{n+1}\phi}{\partial x_0 \dots \partial x_n}(x_0,\dots,x_n). \tag{1.16}$$

Let Π be a cube in R^{n+1} with side length q such that supp $\phi \subset \Pi$. For each $N,$ we divide this cube into N^{n+1} cubes of equal size. For $\nu = 1,\dots,N^{n+1},$ let z^ν denote an interior point of the νth cube whose coordinates are rational numbers. Then we have, for every $\omega \in \Omega,$

$$\int_{R^{n+1}} \psi_i(x_0,\dots,x_n) B_{ij}(x_0,\dots,x_n;\omega)\, dx_0 \cdots dx_n = \lim_{N\to\infty} \frac{q^{n+1}}{N^{n+1}} \sum_{\nu=1}^{N^{n+1}} \langle \omega_j(y),\ F_{ij}(y)\psi_i(z^\nu)\chi_{z^\nu}(y)\rangle. \tag{1.17}$$

But it is easy to see that as a function of $y = (y_0,\dots,y_n),$

$$\frac{q^{n+1}}{N^{n+1}} \sum_{\nu=1}^{N^{n+1}} \psi_i(z^\nu)\chi_{z^\nu}(y)$$

is bounded and compactly supported uniformly in $N,$ and converges to

$$(-1)^k \int_{y_0} \cdots \int_{y_n} \psi_i(x_0,\dots,x_n)\, dx_0 \dots dx_n,$$

as $N \to \infty$, for each y, where k is the number of negative y_i's and

$$\int_{y_i} = \int_{y_i}^{\infty} \qquad \text{for } y_i \geq 0$$

and

$$\int_{y_i} = \int_{-\infty}^{y_i} \qquad \text{for } y_i < 0.$$

Hence, as $N \to \infty$,

$$\frac{q^{n+1}}{N^{n+1}} \sum_{\nu=1}^{N^{n+1}} F_{ij}(y)\psi_i(z^\nu)\chi_{z^\nu}(y) \to (-1)^{n+1} F_{ij}(y)\phi_i(y) \text{ in } L^2(R_y^{n+1}). \tag{1.18}$$

By virtue of (1.7), (1.17), and (1.18), we have (1.14) for almost all ω. □

Next we consider a chaos expansion of the white noise. We follow the construction of an orthonormal basis for $L^2(R^{n+1})$ in [7]. Let $\xi_m(t)$ be the Hermite function of $t \in R$ for $m = 1, 2, \ldots$. Let $\delta^j = (\delta_0^j, \delta_1^j, \ldots, \delta_n^j)$ be the jth multi-index number in some fixed ordering of all $(n+1)$-dimensional multi-indices $\delta = (\delta_0, \ldots, \delta_n)$, each δ_i = a positive integer. This ordering satisfies the property

$$i < j \qquad \text{implies } \delta_0^i + \cdots + \delta_n^i \leq \delta_0^j + \cdots + \delta_n^j.$$

We write

$$\eta_j(x_0, \ldots, x_n) = \xi_{\delta_0^j}(x_0)\xi_{\delta_1^j}(x_1)\cdots\xi_{\delta_n^j}(x_n). \tag{1.19}$$

Then $\{\eta_j\}_{j=1}^{\infty}$ forms an orthonormal basis for $L^2(R^{n+1})$. It follows that for each $\phi \in (C_0^\infty(R^{n+1}))^n$,

$$\begin{aligned} \text{the } i\text{th component of } (F\xi)_\phi(\omega) &= \sum_{j=1}^{n} \langle \omega_j(y),\ F_{ij}(y)\phi_i(y)\rangle \\ &= \lim_{N\to\infty} \sum_{k=1}^{N}\sum_{j=1}^{n} \langle \eta_k(y),\ F_{ij}(y)\phi_i(y)\rangle_{L^2(R_y^{n+1})} \langle \omega_j(y),\ \eta_k(y)\rangle \end{aligned} \tag{1.20}$$

in $L^2(\mathcal{S}', d\mu)$. This will be used to construct a solution in the next section.

We adopt the following definition of a solution to (0.1).

DEFINITION 1.5. *A vector-valued generalized stochastic process $u(t, x_1, \ldots, x_n; \omega)$ is a solution of* (0.1) *in R_+^{n+1} if there is a subset $\Omega \in \mathcal{B}(\mathcal{S}')$ with $\mu(\Omega) = 1$ such that*

$$\ll u,\ L^*\phi \gg = \sum_{i,j=1}^{n} \int_{R_+^{n+1}} (-1)^{n+1} \frac{\partial^{n+1}\phi_i}{\partial t \cdots \partial x_n} B_{ij}(t, x_1, \ldots, x_n; \omega)\, dt \cdots dx_n \tag{1.21}$$

holds for all $\phi \in (C_0^\infty(R_+^{n+1}))^n$, for each $\omega \in \Omega$.

Here, $R_+^{n+1} = \{(t, x) \mid t > 0,\ x \in R^n\}$, $\ll \cdot, \cdot \gg$ is the duality pairing of $(\mathcal{D}'(R_+^{n+1}))^n$ and $(C_0^\infty(R_+^{n+1}))^n$, and L^* is the adjoint of

$$L = \partial_{tt} I - A(t, x, D_x). \tag{1.22}$$

LEMMA 1.6. *Suppose $u(t, x_1, \dots, x_n; \omega)$ is a solution of (0.1) in R_+^{n+1} according to the above definition, i.e., there is a subset $\Omega \in \mathcal{B}(\mathcal{S}')$ with $\mu(\Omega) = 1$ such that for each $\omega \in \Omega$, (1.21) holds for all $\phi \in (C_0^\infty(R_+^{n+1}))^n$. Then, for each $\omega \in \Omega$, $T > 0$ and bounded open subset $\Delta \subset R^n$, we have*

$$\psi(x)u(t, x\,;\,\omega) \in (C([0,T]; H^{-m+1}(R^n)))^n, \tag{1.23}$$

$$\psi(x)u_t(t, x\,;\,\omega) \in (C([0,T]; H^{-m}(R^n)))^n \tag{1.24}$$

for all $\psi \in C_0^\infty(\Delta)$, for some positive integer m.

Proof. Let us set, for $i = 1, \dots, n$,

$$\begin{aligned} &v_i(t, x_1, \dots, x_n; \omega) \\ &\quad = u_i(t, x_1, \dots, x_n; \omega) - \sum_{j=1}^{n} \frac{\partial^n}{\partial x_1 \dots \partial x_n} \int_0^t B_{ij}(s, x_1, \dots, x_n; \omega)\, ds. \end{aligned} \tag{1.25}$$

We fix $\omega \in \Omega$, $T > 0$, and $\Delta \subset R^n$. Let $\eta > 0$ be the number in Theorem 1.2. We note that Theorem 1.2 and Corollary 1.3 are also valid with respect to the reversed time variable. Choose $q > 0$ such that

$$\Delta \subset \{x \in R^n \mid \ |x| < q - T\eta\}. \tag{1.26}$$

Fix any $0 < \epsilon < T/4$, and let $\Psi(t, x) \in C_0^\infty(R_+^{n+1})$ such that

$$\Psi(t, x) = 1 \qquad \text{for } \epsilon \le t \le T + \epsilon,\ |x| \le 2q. \tag{1.27}$$

Since $\Psi v \in (\mathcal{D}'(R_+^{n+1}))^n$ has compact support, there is a positive integer ν such that

$$\Psi v \in (H^{-\nu}(R_+^{n+1}))^n, \tag{1.28}$$

and hence,

$$\Psi v \in (H^{-\nu}(0, \infty\,;\, H^{-\nu}(R^n)))^n. \tag{1.29}$$

In fact, we take $\nu \ge n$ to handle $B_{ij}(t, x_1, \dots, x_n; \omega)$. Next we choose a sequence of functions $\psi_i(t, x) \in C_0^\infty(R_+^{n+1})$, $i = 1, \dots, \nu + 2$, such that $\Psi = 1$ on the support of ψ_1, $\psi_i = 1$ on the support of ψ_{i+1}, for $i = 1, \dots, \nu + 1$, and $\psi_{\nu+2} = 1$, for $2\epsilon \le t \le T$, $|x| \le q$. It is easy to see that

$$\begin{aligned} &\frac{\partial^2}{\partial t^2}(\psi_1 v) \,-\, A(t, x, D_x)(\psi_1 v) \\ &= \psi_1\, A(t, x, D_x) \int_0^t \frac{\partial^n}{\partial x_1 \dots \partial x_n} B(s, x_1, \dots, x_n; \omega)\, ds \\ &\quad + \big(\text{a linear combination of } v \text{ and first order derivatives of } v \\ &\qquad \text{which vanishes outside the support of } \psi_1\big), \end{aligned} \tag{1.30}$$

in the sense of distribution over R_+^{n+1}, where B is R^n-valued, and its ith component is $\sum_{j=1}^n B_{ij}(t, x_1, \dots, x_n; \omega)$.

Since $\psi_1 v = \psi_1 \Psi v \in (H^{-\nu}(0, \infty\,;\, H^{-\nu}(R^n)))^n$, it follows from (1.30)

$$\psi_1 v \in (H^{-\nu+1}(0, \infty\,;\, H^{-\nu-2}(R^n)))^n. \tag{1.31}$$

By repetition of this procedure, we arrive at

$$\psi_{\nu+2}v \in (H^2(0,\infty\,;\,H^{-3\nu-4}(R^n)))^n, \tag{1.32}$$

and hence,

$$\psi_{\nu+2}v \in (C((0,\infty)\,;\,H^{-3\nu-2}(R^n)))^n \cap (C^1((0,\infty)\,;\,H^{-3\nu-4}(R^n)))^n. \tag{1.33}$$

By taking $t = T/2$ as the initial time, we consider the forward and backward Cauchy problem

$$L\theta = \Phi(x)\,A(t,x,D_x)\int_0^t \frac{\partial^n}{\partial x_1\dots\partial x_n}B(s,x_1,\dots,x_n;\omega)\,ds, \tag{1.34}$$

$$\theta(T/2,x) = (\psi_{\nu+2}v)(T/2,x), \qquad \theta_t(T/2,x) = (\psi_{\nu+2}v)_t(T/2,x), \tag{1.35}$$

where $\Phi(x) \in C_0^\infty(R^n)$ with $\Phi(x) = 1$, for $|x| \le q$.

By Theorem 1.1, there is a unique solution

$$\theta \in (C([0,T];H^{-3\nu-3}(R^n)))^n \cap (C^1([0,T];H^{-3\nu-4}(R^n)))^n. \tag{1.36}$$

But, by Corollary 1.3, (1.26), (1.30), (1.34), and (1.35), it holds that

$$\psi(x)v(t,x) = \psi(x)\theta(t,x) \qquad \text{in } [2\epsilon,T]\times\Delta \tag{1.37}$$

for every $\psi(x) \in C_0^\infty(\Delta)$. Now suppose we started out with a smaller number $\tilde\epsilon < \epsilon$. Let $\tilde\nu$ and $\tilde\theta$ correspond to $\tilde\epsilon$ in (1.28) through (1.37). Since $\theta(T/2,x) = \tilde\theta(T/2,x)$ and $\theta_t(T/2,x) = \tilde\theta_t(T/2,x)$ in the open ball $|x| < q$, it again follows from Corollary 1.3 and (1.26) that

$$\psi(x)\theta(t,x) = \psi(x)\tilde\theta(t,x) \qquad \text{in } [0,T]\times\Delta \tag{1.38}$$

for all $\psi \in C_0^\infty(\Delta)$. Meanwhile, we have

$$\psi(x)v(t,x) = \psi(x)\tilde\theta(t,x) \qquad \text{in } [2\tilde\epsilon,T]\times\Delta \tag{1.39}$$

for all $\psi \in C_0^\infty(\Delta)$. By virtue of (1.37), (1.38), and (1.39), we can maintain the same ν for any smaller ϵ and conclude

$$\psi(x)v(t,x) = \psi(x)\theta(t,x) \qquad \text{in } (0,T]\times\Delta \tag{1.40}$$

for every $\psi \in C_0^\infty(\Delta)$. Consequently, (1.23) and (1.24) follow with $m = 3\nu+4$. □

LEMMA 1.7. *In the same setting as above, there is some $\Omega \in \mathcal{B}(\mathcal{S}')$ with $\mu(\Omega) = 1$ such that for each $\omega \in \Omega$, $B_{ij}(0,x;\omega) = 0$ for all $x \in R^n$ and*

$$\begin{aligned} &-\langle u_t(0,x;\omega),\ \phi(0,x)\rangle_\sharp + \langle u(0,x;\omega),\ \phi_t(0,x)\rangle_\sharp \\ &+\int_0^\infty \langle u(t,x;\omega),\ L^*\phi(t,x)\rangle_\sharp\,dt \\ &= \sum_{i,j=1}^n \int_{R_+^{n+1}} (-1)^{n+1}\frac{\partial^{n+1}\phi_i(t,x)}{\partial t\cdots\partial x_n}B_{ij}(t,x_1,\dots,x_n;\omega)\,dt\,dx_1\cdots dx_n \end{aligned} \tag{1.41}$$

for every $\phi(t,x) \in (C_0^\infty(R^{n+1}))^n$, *where* $\langle\cdot\ ,\ \cdot\rangle_\sharp$ *is the duality pairing between* $(\mathcal{D}'(R^n))^n$ *and* $(C_0^\infty(R^n))^n$.

Proof. By the same argument as in the first paragraph of the proof of Lemma 1.4, there is some $\Omega \in \mathcal{B}(\mathcal{S}')$ with $\mu(\Omega) = 1$ such that for each $\omega \in \Omega$, $B_{ij}(0,x;\omega) = 0$, for all $x \in R^n$. By modifying Ω, if necessary, we may assume that for each $\omega \in \Omega$, (1.21) holds for all $\phi \in \big(C_0^\infty(R^{n+1})\big)^n$. Now let us fix $\omega \in \Omega$. Since the tensor product $C_0^\infty(R) \otimes C_0^\infty(R^n)$ is sequentially dense in $C_0^\infty(R^{n+1})$, it is enough to consider ϕ in the form of $\alpha(t)\beta(x)$. Choose any $\alpha(t) \in C_0^\infty(R)$ and $\beta(x) \in (C_0^\infty(R^n))^n$, and choose $T > 0$ such that $\alpha(t) = 0$ for $t \geq T$. We may take Δ and q in (1.26) such that the support of $\beta \subset \Delta$. Since $B_{ij}(t,x;\omega)$ is continuous in $(t,x) \in R^{n+1}$, we can extend the solution of (1.34) and (1.35) to a larger time interval so that

$$\langle u(t,x;\omega),\ \beta(x)\rangle_\sharp\ \in C^1\big([-\epsilon,\, T+\epsilon]\big), \tag{1.42}$$

$$\langle u(t,x;\omega),\ A^*(t,x,D_x)\beta(x)\rangle_\sharp\ \in C^1\big([-\epsilon,\, T+\epsilon]\big) \tag{1.43}$$

for some positive constant ϵ, and

$$\begin{aligned}&\frac{d^2}{dt^2}\langle u*\rho_h,\ \beta\rangle_\sharp(t) - \big(\langle u,\ A^*\beta\rangle_\sharp * \rho_h\big)(t)\\ &= \frac{d}{dt}\sum_{i,j=1}^n \left\langle B_{ij}*\rho_h,\ (-1)^n \frac{\partial^n \beta_i}{\partial x_1 \cdots \partial x_n}\right\rangle_\sharp (t)\end{aligned} \tag{1.44}$$

for each $t \in (-\epsilon/2,\, T+\epsilon/2)$, $0 < h < \epsilon/2$, where $A^*(t,x,D_x)$ is the adjoint of $A(t,x,D_x)$, $\rho_h(t) = \rho(t/h)/h$, with $\rho(t) \in C_0^\infty(R)$ satisfying

$$\rho(t) = \rho(-t) \geq 0; \qquad \rho(t) = 0 \text{ for } |t| \geq 1; \qquad \int_{-\infty}^{\infty} \rho(t)\,dt = 1. \tag{1.45}$$

The convolution is taken with respect to the time variable.

Now we obtain (1.41) by going through the standard procedure: (i) multiply (1.44) by $\alpha(t)$ for $-\epsilon/2 < t < T + \epsilon/2$; (ii) integrate over $[0,T]$; (iii) pass $h \to 0$ with the help of (1.42), (1.43), and the fact $B_{ij}(0,x;\omega) = 0$ for all $x \in R^n$. □

Next we show that the traces of u and u_t with respect to the time variable can be defined as generalized stochastic processes.

LEMMA 1.8. *Let* $u(t,x_1,\dots,x_n;\omega)$ *be a solution of* (0.1) *in* R_+^{n+1}. *Then, for each fixed* $t_0 > 0$, $u(t_0,x_1,\dots,x_n;\omega)$ *and* $u_t(t_0,x_1,\dots,x_n;\omega)$ *are generalized stochastic processes with respect to* $\mathcal{F}(\mathcal{S}')$ *which is the completion of* $\mathcal{B}(\mathcal{S}')$.

Proof. Choose any $\beta(x_1,\dots,x_n) \in (C_0^\infty(R^n))^n$. We have to show that

$$\langle u(t_0,x_1,\dots,x_n;\omega),\ \beta(x_1,\dots,x_n)\rangle_\sharp$$

and

$$\langle u(t_0,x_1,\dots,x_n;\omega),\ \beta(x_1,\dots,x_n)\rangle_\sharp$$

are measurable with respect to $\mathcal{F}(\mathcal{S}')$. Here $\langle\cdot\ ,\ \cdot\rangle_\sharp$ is the duality pairing between $(\mathcal{D}'(R^n))^n$ and $(C_0^\infty(R^n))^n$. Choose a positive number T and an open bounded subset $\Delta \subset R^n$ so that $t_0 < T$ and the support of $\beta \subset \Delta$. Then, by Lemma 1.6, for each fixed $\omega \in \Omega$, where $\Omega \in \mathcal{B}(\mathcal{S}')$ with $\mu(\Omega) = 1$,

$$\langle u(t,x_1,\dots,x_n;\omega),\ \beta(x_1,\dots,x_n)\rangle_\sharp$$

is continuous in $t \in [0,T]$. Consequently,

$$\begin{aligned}(1.46)\qquad &\langle u(t_0,x_1,\dots,x_n;\omega),\ \beta(x_1,\dots,x_n)\rangle_\sharp \\ &= \lim_{h\to 0} \ll u(t,x_1,\dots,x_n;\omega),\ \rho_h(t-t_0)\beta(x_1,\dots,x_n) \gg,\end{aligned}$$

where $\ll \cdot\,,\ \cdot \gg$ is the duality pairing between $(\mathcal{D}'(R^{n+1}_+))^n$ and $(C^\infty_0(R^{n+1}_+))^n$, $\rho_h(t) = \rho(t/h)/h$ and $\rho(t)$ is the same as in (1.45). But, for each sufficiently small h,

$$\ll u(t,x_1,\dots,x_n;\omega),\ \rho_h(t-t_0)\beta(x_1,\dots,x_n) \gg$$

is measurable with respect to $\mathcal{B}(\mathcal{S}')$, because u is a generalized stochastic process with respect to $\mathcal{B}(\mathcal{S}')$. Since (1.46) holds for each $\omega \in \Omega$,

$$\langle u(t_0,x_1,\dots,x_n;\omega),\ \beta(x_1,\dots,x_n)\rangle_\sharp$$

is measurable with respect to $\mathcal{F}(\mathcal{S}')$. By the same argument,

$$\langle u_t(t_0,x_1,\dots,x_n;\omega),\ \beta(x_1,\dots,x_n)\rangle_\sharp$$

is also measurable with respect to $\mathcal{F}(\mathcal{S}')$. □

By virtue of Theorem 1.1, Corollary 1.3, and Lemma 1.6, we have the uniqueness of the solution to (0.1) and (0.2) in the following form.

LEMMA 1.9. *If u_1 and u_2 are solutions of* (0.1) *and* (0.2), *then*

$$(1.47)\qquad u_1 = u_2 \qquad \text{in } R^{n+1}_+$$

for almost all ω.

Let us write with the same η as in Theorem 1.2,

$$(1.48)\qquad \Xi(t_0,x_0;\eta) = \{(t,x) \in R^{n+1} \mid 0 \le t \le t_0,\ |x-x_0| \le \eta(t_0-t)\},$$

and let d be the smallest integer larger than $(n/2)-1$. Our main result is the following.

THEOREM 1.10. *There is a continuous Gaussian process $V(t,x_1,\dots,x_n;\omega)$ with parameter $(t,x_1,\dots,x_n) \in [0,\infty)\times R^n$ such that*

(i) *$L^d V$ is a unique solution of* (0.1) *and* (0.2);

(ii) *$V(t,x_1,\dots,x_n;\omega)$ and $V(\tilde t,\tilde x_1,\dots,\tilde x_n;\omega)$ are independent random variables if $\Xi(t,x_1,\dots,x_n;\eta)$ and $\Xi(\tilde t,\tilde x_1,\dots,\tilde x_n;\eta)$ are disjoint.*

The operator L was defined by (1.22).

2. Proof of the main result. The proof of Theorem 1.10 consists of two basic steps. First, we construct an integral kernel which represents a solution of the deterministic equation:

$$(2.1)\qquad L^d v = f \qquad \text{in } R^{n+1}_+,$$

with zero initial conditions and $f \in L^2_{loc}([0,\infty);(L^2(R^n))^n)$. For this, we use Theorems 1.1, 1.2, and Corollary 1.3. We then obtain approximate solutions by truncating the chaos expansion of the space-time white noise, and prove the convergence to the true solution. Approximation of the white noise is necessary to establish rigorously the existence of a solution according to Definition 1.5.

2.1. Construction of an integral kernel. Consider the initial value problem

$$\begin{cases} Lu^{(1)} = f \qquad \text{in } R_+^{n+1}, \\ u^{(1)}(0,x) = 0, \;\; u_t^{(1)}(0,x) = 0 \quad \text{for } x \in R^n, \end{cases} \tag{2.2}$$

where $f \in L^2_{loc}\big([0,\infty); (L^2(R^n))^n\big)$. Then there is a unique solution

$$u^{(1)} \in C\big([0,\infty); (H^1(R^n))^n\big) \cap C^1\big([0,\infty); (L^2(R^n))^n\big),$$

which satisfies

$$\|u^{(1)}(t,\cdot)\|_{(H^1(R^n))^n} + \|u_t^{(1)}(t,\cdot)\|_{(L^2(R^n))^n} \le C_T \|f\|_{L^2(0,T;(L^2(R^n))^n)} \tag{2.3}$$

for all $0 \le t \le T$. Here and below, C_T stands for generic positive constants depending only on $T > 0$. Next we consider

$$\begin{cases} Lu^{(2)} = u^{(1)} \qquad \text{in } R_+^{n+1}, \\ u^{(2)}(0,x) = 0, \;\; u_t^{(2)}(0,x) = 0 \quad \text{for } x \in R^n. \end{cases} \tag{2.4}$$

We have

$$u^{(2)} \in C\big([0,\infty); (H^2(R^n))^n\big) \cap C^1\big([0,\infty); (H^1(R^n))^n\big),$$

which satisfies

$$\|u^{(2)}(t,\cdot)\|_{(H^2(R^n))^n} + \|u_t^{(2)}(t,\cdot)\|_{(H^1(R^n))^n} \le C_T \|f\|_{L^2(0,T;(L^2(R^n))^n)} \tag{2.5}$$

for all $0 \le t \le T$.

Inductively, we have for $i = 1, \dots, d$,

$$\begin{cases} Lu^{(i+1)} = u^{(i)} \qquad \text{in } R_+^{n+1}, \\ u^{(i+1)}(0,x) = 0, \;\; u_t^{(i+1)}(0,x) = 0 \quad \text{for } x \in R^n. \end{cases} \tag{2.6}$$

It follows that

$$L^{d+1} u^{(d+1)} = f \qquad \text{in } R_+^{n+1} \tag{2.7}$$

and

$$u^{(d+1)} \in C\big([0,\infty); (H^{d+1}(R^n))^n\big) \cap C^1\big([0,\infty); (H^d(R^n))^n\big); \tag{2.8}$$

$$\|u^{(d+1)}(t,\cdot)\|_{(H^{d+1}(R^n))^n} + \|u_t^{(d+1)}(t,\cdot)\|_{(H^d(R^n))^n} \le C_T \|f\|_{L^2(0,T;(L^2(R^n))^n)} \tag{2.9}$$

for all $0 \le t \le T$. Recall that d is the smallest integer larger than $(n/2) - 1$. Let $C_B(R^n)$ be the space of all uniformly bounded continuous functions on R^n. Since $H^{d+1}(R^n) \subset C_B(R^n)$, we find that for each fixed $0 \le t_0 \le T$, $x_0 \in R^n$, the mapping

$$f \mapsto u^{(d+1)}(t_0, x_0) \tag{2.10}$$

is a bounded linear mapping from $(L^2(0,T; L^2(R^n)))^n$ to R^n. Thus there is a matrix function $G = [G_{ij}(t,x;t_0,x_0)]$ with each $G_{ij} \in L^2(0,T; L^2(R^n))$ so that

$$u_i^{(d+1)}(t_0,x_0) = \int_0^T \int_{R^n} \sum_{j=1}^n G_{ij}(t,x;t_0,x_0) f_j(t,x)\, dx\, dt. \tag{2.11}$$

By Corollary 1.3, we find that modification of f outside of $\cup_{0\le t\le t_0}\{(t,\overline{\Gamma_{(t_0,x_0)}(t;\epsilon,\eta)})\}$ does not change the values of $u^{(k)},\ k=1,\dots,d+1,$ in $\cup_{0\le t\le t_0}\{(t,\overline{\Gamma_{(t_0,x_0)}(t;\epsilon,\eta)})\}$. Consequently, it holds that

$$u_i^{(d+1)}(t_0,x_0)=\int_0^{t_0}\int_{\Gamma_{(t_0,x_0)}(t;\epsilon,\eta)}\sum_{j=1}^n G_{ij}(t,x;t_0,x_0)f_j(t,x)\,dx\,dt \tag{2.12}$$

for every $\epsilon>0$. By passing $\epsilon\to 0,$ we arrive at

$$u_i^{(d+1)}(t_0,x_0)=\iint_{\Xi(t_0,x_0;\eta)}\sum_{j=1}^n G_{ij}(t,x;t_0,x_0)f_j(t,x)\,dx\,dt \tag{2.13}$$

which, together with (2.11), implies

$$\text{the support of } G_{ij}(t,x;t_0,x_0)\subset\Xi(t_0,x_0;\eta). \tag{2.14}$$

LEMMA 2.1. *For each $T>0$, the mapping $(t_0,x_0)\mapsto G_{ij}(\,\cdot\,;t_0,x_0)$ is Hölder continuous from $[0,T]\times R^n$ to $L^2([0,T]\times R^n)$, $i,j=1,\dots,n$.*

Proof. Let us write $d+1=(n/2)+\epsilon,$ where $\epsilon=1/2$ if n is odd, and $\epsilon=1$ if n is even. It follows from (2.8) and (2.9)

(2.15)

$$\begin{aligned}
&|u^{(d+1)}(t_0,x_0)-u^{(d+1)}(\tilde t_0,x_0)|\\
&\le C\|u^{(d+1)}(t_0,\cdot)-u^{(d+1)}(\tilde t_0,\cdot)\|_{(H^{(n/2)+(\epsilon/2)}(R^n))^n}\\
&\le C\|u^{(d+1)}(t_0,\cdot)-u^{(d+1)}(\tilde t_0,\cdot)\|^{1-(\epsilon/2)}_{(H^{(n/2)+\epsilon}(R^n))^n}\,\|\int_{\tilde t_0}^{t_0}u_t^{(d+1)}(t,\cdot)\,dt\|^{\epsilon/2}_{(H^{(n/2)+\epsilon-1}(R^n))^n}\\
&\le C_T|t_0-\tilde t_0|^{\epsilon/2}\,\|f\|_{L^2(0,T;(L^2(R^n))^n)}
\end{aligned}$$

for all $t_0,\tilde t_0\in[0,T]$ and all $x_0\in R^n$. In the meantime, by (2.9) and the Sobolev imbedding theorem,

$$|u^{(d+1)}(t_0,x_0)-u^{(d+1)}(t_0,\tilde x_0)|\le C_T|x_0-\tilde x_0|^{\epsilon}\,\|f\|_{L^2(0,T;(L^2(R^n))^n)} \tag{2.16}$$

for all $t_0\in[0,T]$ and all $x_0,\tilde x_0\in R^n$.

This, together with (2.11) and (2.15), yields

$$\begin{aligned}
&\left|\int_0^T\int_{R^n}\sum_{j=1}^n\big(G_{ij}(t,x;t_0,x_0)\ -\ G_{ij}(t,x;\tilde t_0,\tilde x_0)\big)f_j(t,x)\,dx\,dt\right|\\
&\qquad\le C_T(|t_0-\tilde t_0|^{\epsilon/2}+|x_0-\tilde x_0|^{\epsilon})\ \|f\|_{(L^2([0,T]\times R^n))^n},
\end{aligned} \tag{2.17}$$

and thus, for each $i,j=1,\dots,n$,

$$\|G_{ij}(\,\cdot\,;t_0,x_0)-G_{ij}(\,\cdot\,;\tilde t_0,\tilde x_0)\|_{L^2([0,T]\times R^n)}\le C_T(|t_0-\tilde t_0|^{\epsilon/2}+|x_0-\tilde x_0|^{\epsilon}) \tag{2.18}$$

for all $(t_0,x_0),\ (\tilde t_0,\tilde x_0)\in[0,T]\times R^n$. This proves the Hölder continuity of G_{ij}. ∎

2.2. Approximation and convergence. We approximate $F(t,x)$ in (0.1) by

$$F^N(t,x)=\begin{cases}F(t,x) & \text{for } |t|+|x|\le N,\\ 0 & \text{otherwise.}\end{cases} \tag{2.19}$$

Recalling (1.20), we define

$$Q_i^N(t,x;\omega)=\sum_{k=1}^{N}\sum_{j=1}^{n}F_{ij}^N(t,x)\eta_k(t,x)\langle\omega_j,\ \eta_k\rangle. \tag{2.20}$$

Then, for each ω,

$$Q_i^N(t,x;\omega)\in L^2(R_+^{n+1}),\ \ i=1,\dots,n. \tag{2.21}$$

Next we define

$$W_i^N(t,x;\omega)=\int_0^\infty\int_{R^n}\sum_{j=1}^{n}G_{ij}(s,y;t,x)Q_j^N(s,y;\omega)\,dy\,ds. \tag{2.22}$$

Let $W^N=(W_1^N,\dots,W_n^N)$, and $Q^N=(Q_1^N,\dots,Q_n^N)$. Then it holds that for each ω,

$$L^{d+1}W^N=Q^N(t,x;\omega)\qquad\text{in } R_+^{n+1}, \tag{2.23}$$

and, for $j=0,1,\dots,d$,

$$L^jW^N\in C\big([0,\infty);(H^{d+1-j}(R^n))^n\big)\cap C^1\big([0,\infty);(H^{d-j}(R^n))^n\big), \tag{2.24}$$

$$(L^jW^N)(0,x;\omega)=0,\qquad (L^jW^N)_t(0,x;\omega)=0\quad\text{in } R^n. \tag{2.25}$$

Choose any $\mathcal{M}\in\mathcal{B}(\mathcal{S}')$. By virtue of the special structure of W_i^N defined by (2.20) and (2.22), it is apparent that

$$\iint\limits_{R_+^{n+1}}\int_{\mathcal{M}}W^N\cdot\big((L^*)^{d+1}\phi\big)\,d\mu\,dx\,dt=\int_{\mathcal{M}}\iint\limits_{R_+^{n+1}}W^N\cdot\big((L^*)^{d+1}\phi\big)\,dx\,dt\,d\mu \tag{2.26}$$

for all $\phi\in(C_0^\infty(R^{n+1}))^n$. At the same time, we use (2.21), (2.23), (2.24), (2.25), and the same argument as in the proof of Lemma 1.7 to find that

$$\int_{\mathcal{M}}\iint\limits_{R_+^{n+1}}W^N\cdot\big((L^*)^{d+1}\phi\big)\,dx\,dt\,d\mu=\int_{\mathcal{M}}\iint\limits_{R_+^{n+1}}Q^N\cdot\phi\,dx\,dt\,d\mu \tag{2.27}$$

for all $\phi\in(C_0^\infty(R^{n+1}))^n$.

Recalling (1.20), (2.20), and (2.22), we define, for $i=1,\dots,n$,

$$W_i(t,x;\omega)=\sum_{\nu,j=1}^{n}\langle\omega_j(s,y),\ G_{i\nu}(s,y;t,x)F_{\nu j}(s,y)\rangle, \tag{2.28}$$

and write $W = (W_1, \dots, W_n)$. We can use (1.7) in the same way as for (1.20) to find that for each $g \in (L^2(R^{n+1}))^n$,

$$\sum_{k=1}^{N} \sum_{j=1}^{n} \langle \omega_j(z),\, \eta_k(z) \rangle \int_{R^{n+1}} \eta_k(z) g_j(z)\, dz \,\to \sum_{j=1}^{n} \langle \omega_j(z),\, g_j(z) \rangle, \tag{2.29}$$

in $L^2(\mathcal{S}', d\mu)$. The assumption (V), (2.14), (2.18), and (2.29) imply that as $N \to \infty$

$$W_i^N(t,x;\omega) \to W_i(t,x;\omega) \qquad \text{in } L^2(\mathcal{S}', d\mu), \tag{2.30}$$

uniformly in (t,x) of each bounded subset of $[0,\infty) \times R^n$. Hence, we find that for every $\phi \in (C_0^\infty(R^{n+1}))^n$,

$$\iint\limits_{R_+^{n+1}} \int_{\mathcal{M}} W^N \cdot \big((L^*)^{d+1}\phi\big)\, d\mu\, dx\, dt \to \iint\limits_{R_+^{n+1}} \int_{\mathcal{M}} W \cdot \big((L^*)^{d+1}\phi\big)\, d\mu\, dx\, dt, \tag{2.31}$$

as $N \to \infty$.

Meanwhile, for every $\phi \in (C_0^\infty(R^{n+1}))^n$, as $N \to \infty$,

$$\int_{\mathcal{M}} \iint\limits_{R_+^{n+1}} Q^N \cdot \phi\, dx\, dt\, d\mu \to \int_{\mathcal{M}} \sum_{i,j=1}^{n} \langle \omega_j(t,x),\ F_{ij}(t,x)\phi_i(t,x)\pi_+(t) \rangle\, d\mu, \tag{2.32}$$

where

$$\pi_+(t) = \begin{cases} 1 & \text{for } t \ge 0, \\ 0 & \text{otherwise.} \end{cases} \tag{2.33}$$

Thus, for every $\phi \in (C_0^\infty(R^{n+1}))^n$,

$$\iint\limits_{R_+^{n+1}} \int_{\mathcal{M}} W \cdot \big((L^*)^{d+1}\phi\big)\, d\mu\, dx\, dt = \int_{\mathcal{M}} \sum_{i,j=1}^{n} \langle \omega_j,\ F_{ij}\phi_i\pi_+ \rangle\, d\mu. \tag{2.34}$$

Next we will obtain a continuous version of W. Let K be a bounded subset of $[0,\infty) \times R^n$. It follows from (1.7) and (2.18) that for each positive integer m,

$$E\big(\big|W_i(t,x;\omega) - W_i(\tilde{t},\tilde{x};\omega)\big|^{2m}\big) \le C_{m,K} |(t,x) - (\tilde{t},\tilde{x})|^{\epsilon m} \tag{2.35}$$

for all (t,x), $(\tilde{t},\tilde{x}) \in K$, for some positive constant $C_{m,K}$. By a partition of unity and Kolmogorov's theorem, there is a continuous version $V_i(t,x;\omega)$ of $W_i(t,x;\omega)$. We write $V = (V_1, \dots, V_n)$. It is evident that V is a continuous Gaussian process with parameter $(t,x) \in [0,\infty) \times R^n$.

LEMMA 2.2. *For every $\phi \in (C_0^\infty(R^{n+1}))^n$, and every $\mathcal{M} \in \mathcal{B}(\mathcal{S}')$, it holds that*

$$\begin{aligned} &\int_{\mathcal{M}} \iint\limits_{R_+^{n+1}} V \cdot \big((L^*)^{d+1}\phi\big)\, dx\, dt\, d\mu \\ &= \int_{\mathcal{M}} \sum_{i,j=1}^{n} \iint\limits_{R_+^{n+1}} (-1)^{n+1} \frac{\partial^{n+1}\phi_i}{\partial t \partial x_1 \cdots \partial x_n} B_{ij}(t,x;\omega)\, dx\, dt\, d\mu. \end{aligned} \tag{2.36}$$

Proof. Since $E(|W_i(t,x;\cdot)|^2)$ is bounded uniformly in (t,x) of each bounded subset of $[0,\infty)\times R^n$, it is apparent that

$$\iiint_{R_+^{n+1}} \int_{\mathcal{M}} \left|W\cdot\left((L^*)^{d+1}\phi\right)\right| d\mu\,dx\,dt < \infty \tag{2.37}$$

for each $\phi\in(C_0^\infty(R^{n+1}))^n$. In the meantime, $V_i(t,x;\omega)=W_i(t,x;\omega)$ for almost all ω, for each fixed (t,x) so that

$$\iiint_{R_+^{n+1}} \int_{\mathcal{M}} W\cdot\left((L^*)^{d+1}\phi\right) d\mu\,dx\,dt = \iiint_{R_+^{n+1}} \int_{\mathcal{M}} V\cdot\left((L^*)^{d+1}\phi\right) d\mu\,dx\,dt \tag{2.38}$$

and also,

$$\iiint_{R_+^{n+1}} \int_{\mathcal{M}} \left|W\cdot\left((L^*)^{d+1}\phi\right)\right| d\mu\,dx\,dt = \iiint_{R_+^{n+1}} \int_{\mathcal{M}} \left|V\cdot\left((L^*)^{d+1}\phi\right)\right| d\mu\,dx\,dt \tag{2.39}$$

for every $\phi\in(C_0^\infty(R^{n+1}))^n$. Furthermore, $V_i(t,x;\omega)$ is continuous in (t,x) for each ω, and measurable with respect to $\mathcal{B}(\mathcal{S}')$ for each (t,x). Hence, V_i is measurable with respect to the product σ-algebra. By Fubini's theorem, we have

$$\iiint_{R_+^{n+1}} \int_{\mathcal{M}} V\cdot\left((L^*)^{d+1}\phi\right) d\mu\,dx\,dt = \int_{\mathcal{M}} \iiint_{R_+^{n+1}} V\cdot\left((L^*)^{d+1}\phi\right) dx\,dt\,d\mu. \tag{2.40}$$

Meanwhile, it follows from Lemma 1.4 that for every $\phi\in(C_0^\infty(R^{n+1}))^n$,

$$\begin{aligned}&\int_{\mathcal{M}} \sum_{i,j=1}^n <\omega_j,\ F_{ij}\phi_i>\ d\mu\\ &= \int_{\mathcal{M}} \sum_{i,j=1}^n \iiint_{R^{n+1}} (-1)^{n+1}\frac{\partial^{n+1}\phi_i}{\partial t\partial x_1\dots\partial x_n} B_{ij}(t,x_1,\dots,x_n;\omega)\,dx\,dt\,d\mu.\end{aligned} \tag{2.41}$$

Let us choose $\pi(t)\in C^\infty(R)$ with $\pi(t)=1$ for $t\ge 0$, and $\pi(t)=0$ for $t\le -1$, and write

$$\pi^h(t)=\pi(t/h) \qquad \text{for } h>0. \tag{2.42}$$

Then (2.41) is valid with ϕ_i replaced by $\phi_i\pi^h$, and it is easy to see that as $h\to 0$,

$$\sum_{i,j=1}^n \langle\omega_j,\ F_{ij}\phi_i\pi^h\rangle \ \to \sum_{i,j=1}^n \langle\omega_j,\ F_{ij}\phi_i\pi_+\rangle \qquad \text{in } L^2(\mathcal{S}',d\mu). \tag{2.43}$$

Since $B_{ij}(t,x;\omega)$ is continuous in (t,x) for each ω, it follows from (1.15) and Fatou's lemma that for each bounded subset $K\subset[0,\infty)\times R^n$,

$$E\left(\left|B_{ij}(t,x;\cdot)\right|^2\right)\le C_K \tag{2.44}$$

for all $(t,x) \in K$, for some positive constant C_K. We use this for the following procedure:

$$
\begin{aligned}
(2.45)\qquad & \lim_{h\to 0} \int_{\mathcal{M}} \iint_{R^{n+1}} \frac{\partial^n \phi_i(t,x)}{\partial x_1 \cdots \partial x_n} \frac{1}{h} \pi'(t/h)\, B_{ij}(t,x;\omega)\, dx\, dt\, d\mu \\
&= \lim_{h\to 0} \int_{-1}^{0} \int_{|x|\le q} \int_{\mathcal{M}} \frac{\partial^n \phi_i}{\partial x_1 \cdots \partial x_n}(hs,x)\, \pi'(s)\, B_{ij}(hs,x;\omega)\, d\mu\, dx\, ds \\
&= \int_{-1}^{0} \int_{|x|\le q} \left(\lim_{h\to 0} \int_{\mathcal{M}} \frac{\partial^n \phi_i}{\partial x_1 \cdots \partial x_n}(hs,x)\, \pi'(s)\, B_{ij}(hs,x;\omega)\, d\mu \right) dx\, ds \\
&= 0.
\end{aligned}
$$

Here the second equality is justified by the fact that

$$
\int_{\mathcal{M}} \frac{\partial^n \phi_i}{\partial x_1 \cdots \partial x_n}(hs,x)\, \pi'(s)\, B_{ij}(hs,x;\omega)\, d\mu
$$

is bounded uniformly in $0 \le h \le 1$, $-1 \le s \le 0$, and $|x| \le q$, where q is a positive number such that $\phi_i(t,x) = 0$ for $|x| \ge q$. For the last equality, we argue as follows. First, for each fixed s and x, as $h \to 0$,

$$
(2.46)\qquad < \omega_j(y),\ F_{ij}(y) \chi_{(hs,x_1,\dots,x_n)}(y) > \to 0 \quad \text{in } L^2(\mathcal{S}', d\mu).
$$

Next, since $B_{ij}(t,x;\omega)$ is a continuous version of the process defined by (1.10), we have for each fixed s and x,

$$
\begin{aligned}
(2.47)\qquad \lim_{h\to 0} \int_{\mathcal{M}} B_{ij}(hs,x;\omega)\, d\mu &= \lim_{h\to 0} \int_{\mathcal{M}} < \omega_j(y),\ F_{ij}(y) \chi_{(hs,x_1,\dots,x_n)}(y) >\, d\mu \\
&= 0.
\end{aligned}
$$

Hence, the last equality follows.

Similarly, we also have

$$
\begin{aligned}
(2.48)\qquad & \lim_{h\to 0} \int_{\mathcal{M}} \iint_{R^{n+1}} \frac{\partial^{n+1} \phi_i(t,x)}{\partial t \cdots \partial x_n}\, \pi^h(t,x) B_{ij}(t,x;\omega)\, dx\, dt\, d\mu \\
&= \int_{\mathcal{M}} \iint_{R_+^{n+1}} \frac{\partial^{n+1} \phi_i(t,x)}{\partial t \cdots \partial x_n}\, B_{ij}(t,x;\omega)\, dx\, dt\, d\mu.
\end{aligned}
$$

Combining (2.43), (2.45), and (2.48), we finally arrive at

$$
\begin{aligned}
(2.49)\qquad & \int_{\mathcal{M}} \sum_{i,j=1}^{n} \langle \omega_j,\ F_{ij} \phi_i \pi_+ \rangle\, d\mu \\
&= \int_{\mathcal{M}} \sum_{i,j=1}^{n} \iint_{R_+^{n+1}} (-1)^{n+1} \frac{\partial^{n+1} \phi_i}{\partial t \partial x_1 \cdots \partial x_n} B_{ij}(t,x;\omega)\, dx\, dt\, d\mu
\end{aligned}
$$

for all $\phi \in (C_0^\infty(R^{n+1}))^n$. By means of (2.34), (2.38), (2.40), and (2.49), we derive (2.36) and conclude the proof of Lemma 2.2. □

We now show that $L^d V(t,x;\omega)$ is a solution of (0.1) according to Definition 1.5. (2.36) implies that for each $\phi \in (C_0^\infty(R_+^{n+1}))^n$,

$$\ll V,\ (L^*)^{d+1}\phi \gg = \sum_{i,j=1}^{n} \iint_{R_+^{n+1}} (-1)^{n+1} \frac{\partial^{n+1}\phi_i}{\partial t \cdots \partial x_n} B_{ij}(t,x;\omega)\, dx\, dt \tag{2.50}$$

holds for almost all ω. Let K be a compact subset of R_+^{n+1}. Then $C_0^\infty(K)$ is a separable Fréchet space. Hence, there is a countable dense subset $\{\phi^{(\nu)}\}_{\nu=1}^\infty \subset \big(C_0^\infty(K)\big)^n$. For each $\phi^{(\nu)}$, there is $\Omega_\nu \in \mathcal{B}(\mathcal{S}')$ such that $\mu(\Omega_\nu) = 1$, and, for all $\omega \in \Omega_\nu$,

$$\ll V,\ (L^*)^{d+1}\phi^{(\nu)} \gg = \sum_{i,j=1}^{n} \iint_{R_+^{n+1}} (-1)^{n+1} \frac{\partial^{n+1}\phi_i^{(\nu)}}{\partial t \cdots \partial x_n} B_{ij}(t,x;\omega)\, dx\, dt \tag{2.51}$$

holds. Let $\Omega_K = \cap_{\nu=1}^\infty \Omega_\nu$. Then, $\mu(\Omega_K) = 1$, and for all $\omega \in \Omega_K$, (2.51) holds for every $\phi \in \big(C_0^\infty(K)\big)^n$. Since R_+^{n+1} is a countable union of compact subsets, there is $\Omega \in \mathcal{B}(\mathcal{S}')$ such that $\mu(\Omega) = 1$, and for every $\omega \in \Omega$, (2.50) holds for all $\phi \in (C_0^\infty(R_+^{n+1}))^n$. Hence, $L^d V$ is a solution of (0.1) in R_+^{n+1}.

Next we will show that $L^d V$ satisfies the initial conditions (0.2). Choose any $\gamma_1(x),\ \gamma_2(x) \in (C_0^\infty(R^n))^n$, and consider the Cauchy problem:

$$L^*\psi = 0 \qquad \text{in } (-\infty,\infty) \times R^n, \tag{2.52}$$

$$\psi(0,x) = \gamma_1(x), \qquad \psi_t(0,x) = \gamma_2(x) \quad \text{in } R^n. \tag{2.53}$$

Choose a function $\zeta(t) \in C_0^\infty(R)$ such that $\zeta(t) = 1$ for $|t| \le 1$, and $\zeta(t) = 0$ for $|t| \ge 2$. Let us set $\sigma(t,x) = \zeta(t)\psi(t,x)$. The solution of the above Cauchy problem satisfies the property of a domain of dependence, and consequently, $\sigma(t,x) \in (C_0^\infty(R^{n+1}))^n$. Furthermore, it is easy to see

$$L^*\sigma = 2\zeta_t\, \psi_t + \zeta_{tt}\psi \in (C_0^\infty(R^{n+1}))^n, \tag{2.54}$$

where the right-hand side vanishes for $|t| \le 1$. In the meantime, by the same argument as above, we can infer from (2.36) that there is some $\Omega \in \mathcal{B}(\mathcal{S}')$ with $\mu(\Omega) = 1$ such that for each $\omega \in \Omega$,

$$\iint_{R_+^{n+1}} V \cdot \big((L^*)^{d+1}\phi\big)\, dx\, dt = \sum_{i,j=1}^{n} \iint_{R_+^{n+1}} (-1)^{n+1} \frac{\partial^{n+1}\phi_i}{\partial t \cdots \partial x_n} B_{ij}(t,x;\omega)\, dx\, dt \tag{2.55}$$

holds for all $\phi \in (C_0^\infty(R^{n+1}))^n$. With $\phi = \sigma$, the left-hand side can be written as

$$\iint_{R_+^{n+1}} V \cdot \big((L^*)^d L^*\sigma\big)\, dx\, dt = \int_0^\infty \langle L^d V,\ L^*\sigma\rangle_\sharp\, dt, \tag{2.56}$$

because $L^*\sigma \in \big(C_0^\infty(R_+^{n+1})\big)^n$ for $t \ge 0$. Thus (2.55) yields

$$\begin{aligned} &\int_0^\infty \langle L^d V(t,x;\omega),\ L^*\sigma(t,x)\rangle_\sharp\, dt \\ &\quad = \sum_{i,j=1}^{n} \iint_{R_+^{n+1}} (-1)^{n+1} \frac{\partial^{n+1}\sigma_i(t,x)}{\partial t \cdots \partial x_n} B_{ij}(t,x;\omega)\, dx\, dt \end{aligned} \tag{2.57}$$

for all $\omega \in \Omega$. On the other hand, Lemma 1.7 implies that

$$
\begin{aligned}
(2.58) \qquad & -\langle (L^d V)_t(0,x;\omega),\ \gamma_1(x)\rangle_\sharp + \langle L^d V(0,x;\omega),\ \gamma_2(x)\rangle_\sharp \\
& + \int_0^\infty \langle L^d V(t,x;\omega),\ L^*\sigma(t,x)\rangle_\sharp \, dt \\
& = \sum_{i,j=1}^{n} \iint_{R_+^{n+1}} (-1)^{n+1} \frac{\partial^{n+1}\sigma_i(t,x)}{\partial t \cdots \partial x_n} B_{ij}(t,x;\omega)\, dx\, dt
\end{aligned}
$$

for all $\omega \in \Omega$, after modification of Ω if necessary. But we note that this Ω is independent of γ_1 and γ_2. By comparing (2.57) and (2.58), we conclude that

$$
L^d V(0,x;\omega) = 0, \qquad (L^d V)_t(0,x;\omega) = 0 \quad \text{in } R^n
$$

for each $\omega \in \Omega$, because γ_1 and γ_2 were chosen arbitrarily. The uniqueness is given by Lemma 1.9

Now the statement (i) of Theorem 1.10 has been established. The statement (ii) follows easily from (1.6), (2.14), and (2.28).

Final remark. Our goal is to obtain a solution in the form (0.6). If one is interested only in the existence of weak solutions for almost all ω, there is a somewhat direct approach. It follows from Lemma 1.4 that $F\xi \in (H^{-1}_{loc}(R; H^{-n}_{loc}(R^n)))^n$ for almost all ω. Since Theorem 1.1 cannot be applied directly for this function class, it requires some work to obtain a weak solution. This involves localization in the space variables and handling low regularity in the time variable as in the proof of Lemma 1.6. It also needs some extra work to show that the resulting weak solution is a generalized stochastic process. Meanwhile, the structure of such a solution is intractable. The reward for our special procedure for the existence of solution is twofold. First, the representation formula shows globally uniform structure of the solution as a generalized stochastic process. Second, our procedure shows that the solution can be approximated by a sequence of ordinary stochastic processes.

REFERENCES

[1] E.M. Cabaña, *The vibrating string forced by white noise*, Z. Wahr scheinlich Keits theorie und Verw, Gebiete, 15 (1970), pp. 111–130.

[2] R. Courant and D. Hilbert, *Methods of Mathematical Physics*, Vol. 2, Interscience Publishers, New York, London, Sydney, 1962.

[3] G. Da Prato and J. Zabczyk, *Stochastic Equations in Infinite Dimensions*, Cambridge University Press, Cambridge, UK, 1992.

[4] R.C. Dalang and N.E. Frangos, *The stochastic wave equation in two spatial dimensions*, Ann. Probab., 26 (1998), pp. 187–212.

[5] G.F.D. Duff, *The Cauchy problem for elastic waves in an anisotropic medium*, Philos. Trans. Roy. Soc. London Ser. A, 252 (1960), pp. 249–273.

[6] B. Gaveau, *The Cauchy problem for the stochastic wave equation*, Bull. Sci. Math., 119 (1995), pp. 381–407.

[7] H. Holden, B. Øksendal, J. Ubøe, T. Zhang, *Stochastic Partial Differential Equations: A Modeling, White Noise Functional Approach*, Birkhäuser, Boston, 1996.

[8] I. Karatzas and S.E. Shreve, *Brownian Motion and Stochastic Calculus*, 2nd ed., Springer-Verlag, New York, 1991.

[9] J.U. Kim, *On the local regularity of solutions in linear viscoelasticity of several space dimensions*, Trans. Amer. Math. Soc., 346 (1994), pp. 359–398.

[10] M. Marcus and V. Mizel, *Stochastic hyperbolic systems and the wave equation*, Stochastics and Stochastics Reports, 36 (1991), pp. 225–244.

[11] C. Mueller, *Long time existence for the wave equation with a noise term*, Ann. Probab., 25 (1997), pp. 133–151.

[12] D. Nualart, *The Malliavin Calculus and Related Topics*, Springer-Verlag, New York, 1995.
[13] E. Orsingher, *Damped vibrations excited by white noise*, Adv. in Appl. Probab., 16 (1984), pp. 562–584.
[14] M. Riesz, *L'intégrale de Riemann-Liouville et le problème de Cauchy*, Acta Math., 81 (1949), pp. 1–223.
[15] J.B. Walsh, *An Introduction to Stochastic Partial Differential Equations*, Lecture Notes in Math. 1180, Springer, Berlin, New York, 1986.

SIAM J. MATH. ANAL.
Vol. 32, No. 2, pp. 323–359

TRAVELING TWO AND THREE DIMENSIONAL CAPILLARY GRAVITY WATER WAVES*

WALTER CRAIG† AND DAVID P. NICHOLLS‡

Abstract. The main results of this paper are existence theorems for traveling gravity and capillary gravity water waves in two dimensions, and capillary gravity water waves in three dimensions, for any periodic fundamental domain. This is a problem in bifurcation theory, yielding curves in the two dimensional case and bifurcation surfaces in the three dimensional case. In order to address the presence of resonances, the proof is based on a variational formulation and a topological argument, which is related to the resonant Lyapunov center theorem.

Key words. water waves, bifurcation theory, traveling waves

AMS subject classifications. 35; 76

PII. S0036141099354181

1. Introduction.

Nonlinear periodic traveling waves on the free surface of an ideal fluid tend to form hexagonal patterns. This phenomenon is the focus of a number of recent papers on the subject of water waves, and it is the topic of the present article. In previous work, various approximations to the evolution equations for free surface waves are used, in particular the KP system by J. Hammack, N. Scheffner, and H. Segur [11], and J. Hammack, D. McCallister, N. Scheffner, and H. Segur [12], and alternatively with certain formal shallow water expansions of the Euler equations by P. Milewski and J.B. Keller [16]. A natural question is whether similar patterns can be shown to occur in solutions of the full Euler equations themselves. This is the focus of a series of papers by the present authors. In [18] and in [20] we report on hexagonal wave patterns and other phenomena in numerical computations of solutions, which are shown to satisfy spectral criteria for numerical convergence to solutions of Euler's equations. In the present article we describe rigorous existence results for periodic traveling wave solutions in free surfaces. The goal is to prove the existence of nontrivial traveling wave solutions to the water wave problem for gravity and capillary gravity waves in two and three dimensions. In two dimensions this is proven for both gravity and capillary gravity water waves, constituting a new and relatively straightforward approach to the theorems of T. Levi-Civita and D. Struik. In three dimensions we prove the existence of traveling capillary gravity water waves. However, the problem of gravity waves in three dimensions exhibits the phenomena of small divisors, and it remains open. The theorem that we prove is given below.

THEOREM 1.1. *For any spatial period there exist nontrivial periodic traveling wave solutions of the water wave problem in two dimensions for gravity and capillary gravity waves, both in deep water and in water of finite depth. In three dimensions, for any periodic fundamental domain of* $\mathbb{R}^2$ *there exist nontrivial periodic traveling wave solutions of the water wave problem for capillary gravity waves in infinite depth*

*Received by the editors March 26, 1999; accepted for publication (in revised form) October 26, 1999; published electronically July 5, 2000.

http://www.siam.org/journals/sima/32-2/35418.html

†Department of Mathematics, Brown University, Providence, RI 02912 (craigw@math.brown.edu). This author was supported by the NSF under grant DMS 9706273.

‡Department of Mathematics, University of Minnesota, Minneapolis, MN 55455 (nicholls@math.umn.edu). This author was partially supported by the Division of Applied Mathematics and the Department of Mathematics at Brown University.

water and for water of finite depth.

We show that in fact there are many small amplitude traveling waves bifurcating from uniform flow. A solution of the two dimensional problem is of course also a three dimensional one which is constant in one independent variable. In contrast, for the three dimensional problem, the solutions asserted by this theorem explicitly give fully three dimensional wave patterns, as we will see below.

The early rigorous theorems on traveling water waves concern the two dimensional gravity wave problem, by T. Levi-Civita (1925, [15]) in an infinitely deep fluid, and with similar complex variable techniques by D. Struik (1926, [27]) for a fluid of finite depth. When surface tension effects are included, the analogous two dimensional traveling wave problem was addressed by E. Zeidler [31] and H. Beckert and E. Zeidler [2] for large coefficient of surface tension, and J.T. Beale [1], who also allowed for small surface tension under conditions of nonresonance. Later J. Reeder and M. Shinbrot reproduced these results [24], and M. Jones and J. Toland [13], [14] dealt specifically with the occurrence of resonance.

The only papers of which we are aware[1] that give a rigorous analysis of the three dimensional water wave problem are by J. Reeder and M. Shinbrot [23] and T.Y. Sun [28], in which they prove an existence theorem for periodic traveling capillary gravity waves, whose fundamental domain is a "symmetric diamond." In both papers the authors remark that the three dimensional gravity wave problem is an example of a small denominator problem. The proof in [23] applies to most choices of the parameters of the problem (the spatial periods, the acceleration of gravity g, the fluid depth h, and the Bond number σ/gh^2). However, they must explicitly avoid cases in which a certain linearized operator has a zero eigenvalue of higher multiplicity. These are situations where more than the generic one or two solutions of the linearized equation have the same phase velocity. It is easily shown that these exceptional cases are common, and that they accumulate in parameter space at zero Bond number. The paper [28] also is restricted in the same way, but is in some ways more general as it also contains results for the case of a pressure disturbance applied to the free surface traveling at the phase velocity. The paper [23] also reproves the two dimensional theorems for gravity and for capillary gravity free surface waves, again with the proviso that the parameters are chosen so that a higher multiplicity null eigenvalue is avoided. The two dimensional case without such proviso appears in [13], [14]. With regard to the three dimensional water wave problem without surface tension, see P. Plotnikov [21].

In this paper our proof is based on a reformulation of the capillary gravity wave problem involving surface integrals, and a variational argument which is similar to that used by A. Weinstein [29] and J. Moser [17] in their work on the resonant Lyapunov center theorem. A higher dimensional null space in the free surface problem compares formally with a case of resonance near an elliptic stationary point of a Hamiltonian system. Furthermore, the water wave problem is Hamiltonian in our chosen coordinates, a fact due initially to V.E. Zakharov [30], and this plays a role in a counting argument via an equivariant cohomological index for a lower bound on the number of distinct solutions. Our approach is bifurcation theoretic, with the novelty that for the three dimensional problem the basic parameter is the two dimensional phase velocity, and solutions occur correspondingly in bifurcation surfaces rather than curves.

The organization of the paper is as follows. In section 2 we introduce the

[1]We have, however, recently received a preprint of the work of M.D. Groves and A. Mielke [10] on three dimensional capillary gravity waves in channels.

Dirichlet–Neumann operator, reformulate the water wave problem in a set of coordinates involving surface integrals which we have found useful [8], [19], and outline the method of Lyapunov–Schmidt. In section 3 we solve the first Lyapunov–Schmidt equation via the implicit function theorem, and in section 4 we pose the problem as a variational problem for the extrema of an action integral. We then reduce the variational problem to a finite dimensional one using the results of section 3, and analyze its set of solutions. Finally, in section 5 we prove the analyticity of the Dirichlet–Neumann operator in appropriate function spaces. In many ways this analysis is the heart of the proof. The method is very similar to the one of W. Craig, U. Schanz, and C. Sulem [7] and D. Nicholls [19]; however, we modify it in a nontrivial way to yield the particular estimates which we require.

2. Formulation of the water wave problem. The water wave problem describes the evolution of an ideal fluid with free surface under the effects of gravity (gravity waves) or gravity and surface tension (capillary gravity waves). An ideal fluid is one which is inviscid, incompressible, and irrotational. To begin we will consider an n dimensional fluid, meaning $n-1$ horizontal dimensions and one vertical dimension, and specialize to two and three dimensions later.

2.1. Classical equations and a surface integral formulation. Consider a fluid region given by

$$S_\eta = \{(x,y) \in \mathbb{R}^{n-1} \times \mathbb{R} \mid -h \leq y \leq \eta(x,t)\}, \tag{2.1}$$

where $\eta(x,t)$ is the free surface, and $0 < h \leq +\infty$. It is well known for an ideal fluid that inside the fluid domain S_η the fluid velocity $\mathbf{u}$ can be expressed as

$$\mathbf{u} = \nabla\varphi, \tag{2.2}$$

since the fluid is irrotational, and further

$$\operatorname{div}[\mathbf{u}] = \operatorname{div}[\nabla\varphi] = \Delta\varphi = 0, \tag{2.3}$$

since the fluid is incompressible. For the problem of water of finite depth the boundary conditions are given by

$$\partial_y\varphi = 0 \qquad \text{at } y = -h, \tag{2.4a}$$

$$\partial_t\eta + \nabla_x\varphi \cdot \nabla_x\eta - \partial_y\varphi = 0 \qquad \text{at } y = \eta(x,t), \tag{2.4b}$$

$$\partial_t\varphi + \tfrac{1}{2}|\nabla\varphi|^2 + g\eta - \sigma\operatorname{div}\left[\frac{\nabla_x\eta}{\sqrt{1+|\nabla_x\eta|^2}}\right] = 0 \qquad \text{at } y = \eta(x,t)\ , \tag{2.4c}$$

where in appropriate units the density of the fluid is taken to be one, g is the acceleration of gravity, and σ/gh^2 is the Bond number of the interface. To consider water of infinite depth we can replace (2.4a) with the following condition:

$$\partial_y\varphi \to 0 \qquad \text{as } y \to -\infty. \tag{2.4d}$$

The horizontal boundary conditions will be periodic, which is to say that in two dimensions our surface is parameterized by a circle, and in three dimensions by a torus; both of these are determined by a lattice $\Gamma \subseteq \mathbb{R}^{n-1}$, generated by a nonsingular matrix $A \in \mathbb{R}^{(n-1)\times(n-1)}$ acting on vectors $j \in \mathbb{Z}^{n-1}$, which in turn has a conjugate

lattice $\Gamma' \subseteq \mathbb{R}^{n-1}$ generated by the matrix $2\pi(A^T)^{-1}$. We recall that a well-behaved function f on the torus $T(\Gamma) = \mathbb{R}^{n-1}/\Gamma$ can be written in terms of its Fourier series

$$f(x) = \sum_{k \in \Gamma'} \hat{f}(k) e^{ik \cdot x}. \tag{2.5}$$

The classical formulation of the water wave problem is given in (2.3), (2.4a) (or (2.4d)), (2.4b), (2.4c) with periodic boundary conditions. We will follow [8], [25], [19], [18] and recast these equations in a surface integral formulation of the water wave problem. We begin with the observation that when the free surface $\eta(x,t)$ and Dirichlet data at the free surface $\xi(x,t) = \varphi(x, \eta(x,t), t)$ are specified, we can in principle solve the full problem, since φ satisfies Laplace's equation with appropriate boundary conditions. In this way the water wave problem is reduced from one posed inside the entire fluid to one posed at the free surface alone. V.E. Zakharov [30] noted this and also made the elegant statement that the surface variables η and ξ are canonically conjugate variables with which one may formulate the water wave problem as a Hamiltonian system, with Hamiltonian

$$H = \int_{T(\Gamma)} \int_{-h}^{\eta} \tfrac{1}{2} |\nabla \varphi|^2 \, dy + \tfrac{1}{2} g \eta^2 + \sigma \left(\sqrt{1 + |\nabla_x \eta|^2} - 1 \right) dx. \tag{2.6}$$

Of course, the full dependence on η and ξ is complicated and not explicit. However, W. Craig and C. Sulem [8] found a more convenient representation of the Hamiltonian by noting that the term $\int_{T(\Gamma)} \int_{-h}^{\eta} |\nabla \varphi|^2 \, dy \, dx$ is a quadratic form in the quantity ξ involving the Dirichlet–Neumann operator.

DEFINITION 2.1. *The Dirichlet–Neumann operator of the free surface is defined by*

$$G(\eta)\, \xi = \nabla \varphi|_{y=\eta} \cdot N_\eta, \tag{2.7}$$

where the potential function φ satisfies (2.3), (2.4a), *$\varphi(x, \eta(x)) = \xi(x)$, and $N_\eta = (-\nabla_x \eta, 1)^T$ is a (nonnormalized) exterior normal.*

This operator is the central tool in our analysis and a deep understanding of its properties is key to the boundary integral formulation. For our purposes its analyticity in appropriate function spaces is the relevant property which we will state and prove in section 5. Using the Dirichlet–Neumann operator and the divergence theorem, the Hamiltonian for the water wave problem can be written as

$$H = \int_{T(\Gamma)} \tfrac{1}{2} \xi (G(\eta)\, \xi) + \tfrac{1}{2} g \eta^2 + \sigma \left(\sqrt{1 + |\nabla_x \eta|^2} - 1 \right) dx. \tag{2.8}$$

By taking the appropriate variations and interpreting them in the correct fashion, or simply taking (2.4b), (2.4c) and substituting in the Dirichlet–Neumann operator in the appropriate way, one arrives at the following formulation of the water wave problem [5], [7]:

$$\partial_t \eta = G(\eta)\, \xi, \tag{2.9a}$$

$$\begin{aligned} \partial_t \xi = -g\eta - \frac{1}{2\left(1 + |\nabla_x \eta|^2\right)} \Big[& |\nabla_x \xi|^2 - (G(\eta)\, \xi)^2 \\ & - 2(G(\eta)\, \xi) \nabla_x \xi \cdot \nabla_x \eta + |\nabla_x \xi|^2 |\nabla_x \eta|^2 - (\nabla_x \xi \cdot \nabla_x \eta)^2 \Big] \end{aligned} \tag{2.9b}$$

$$+\sigma \mathrm{div}\left[\frac{\nabla_x \eta}{\sqrt{1+|\nabla_x \eta|^2}}\right].$$

2.2. A variational principle and the method of Lyapunov–Schmidt. There is a variational formulation, related to the principle of stationary action, whose Euler–Lagrange equations give traveling wave solutions of (2.9). Given the lattice $\Gamma \subseteq \mathbb{R}^{n-1}$, consider a class of mappings from the torus $T(\Gamma) = \mathbb{R}^{n-1}_x/\Gamma$ to a phase space $X = \{u = (\eta(x), \xi(x))^T\}$ (the topology will be specified later). For these mappings we define $n-1$ many *action* functionals

$$I_j(\eta(x), \xi(x)) = \int_{T(\Gamma)} \eta(x) \partial_{x_j} \xi(x)\, dx \tag{2.10}$$

and as well the (averaged) Hamiltonian $H(\eta(x), \xi(x))$ given in (2.8). The formal variational principle is to fix the values of these $n-1$ actions $I_j = a_j$, and consider extremal points of the Hamiltonian H. If such points existed and were sufficiently regular, they would satisfy the Euler–Lagrange equations

$$\delta H = \sum_{j=1}^{n-1} c_j \delta I_j\ , \tag{2.11}$$

which are traveling wave solutions of system (2.9) with Lagrange multiplier the phase velocity $c = (c_1, \ldots, c_{n-1})$. There is a torus action on the phase space X given by $T_\alpha(\eta(x), \xi(x)) = (\eta(x+\alpha), \xi(x+\alpha))$, under which I_j and H are invariant, and in this way every critical point is in fact a member of a torus of critical points in X. These considerations are, however, purely formal, the level sets of I_j are by no means compact, and without further analytic information about the functionals I_j and H this procedure does not give rise to actual solutions of the problem, even in finite dimensional settings such as the Lyapunov center theorem. Nonetheless, a reduction that is related to the method of Lyapunov–Schmidt gives rise to a related finite dimensional variational problem, also invariant with respect to the torus action, whose critical points give traveling wave solutions to (2.9). This procedure is analogous to the resonant Lyapunov center theorem and the reduction by J. Moser [17] of the variational problem to a finite dimensional one which is invariant under a circle action.

In order to set up the problem of traveling surface waves, introduce a frame of reference moving with velocity c, and change variables in (2.9). This leads to the definition of the functions

$$\begin{aligned} F_1(\eta, \xi, c) = {} & g\eta - c \cdot \nabla_x \xi + \frac{1}{2\left(1+|\nabla_x \eta|^2\right)} \Big[|\nabla_x \xi|^2 - (G(\eta)\ \xi)^2 \\ & - 2(G(\eta)\ \xi) \nabla_x \xi \cdot \nabla_x \eta + |\nabla_x \xi|^2 |\nabla_x \eta|^2 - (\nabla_x \xi \cdot \nabla_x \eta)^2 \Big] \\ & - \sigma \mathrm{div}\left[\frac{\nabla_x \eta}{\sqrt{1+|\nabla_x \eta|^2}}\right], \end{aligned} \tag{2.12a}$$

$$F_2(\eta, \xi, c) = c \cdot \nabla_x \eta + G(\eta)\ \xi\ , \tag{2.12b}$$

with which we will abbreviate the problem of traveling waves for the system (2.9) as $F(\eta, \xi, c) = 0$. Let $u = (\eta, \xi)^T$ and $F : X \times C \to Y$, where the Banach spaces X, C,

and Y are to be specified later. A trivial branch of solutions to this problem is given by $(0, c)$ for all c, and from this trivial branch we produce a nontrivial bifurcation branch of solutions. We will use the implicit function theorem and hence must understand the linearization of F about the trivial solution $\partial_u F(0, c) \equiv A(c)$. If $A(c)$ from X to Y is boundedly invertible for some parameter c, then we may refer to the implicit function theorem to obtain a unique solution in a neighborhood of c, namely, the trivial one. The possible bifurcation points are those values of c for which $A(c)$ has a zero eigenvalue. Using the decomposition of Lyapunov–Schmidt, let c_0 be a parameter value for which $A(c_0)$ has a nontrivial null space, and let

$$X_1 = \text{null}(A(c_0)), \tag{2.13a}$$

$$Y_1 = \text{range}(A(c_0)), \tag{2.13b}$$

and X_2 and Y_2 to be such that $X = X_1 \oplus X_2$, $Y = Y_1 \oplus Y_2$. Let P be the orthogonal projection of Y onto Y_1 and $Q = I - P$. The method of Lyapunov–Schmidt replaces the problem $F(u, c) = 0$ by the equivalent pair of equations

$$PF(v + w, c) = 0, \tag{2.14a}$$

$$QF(v + w, c) = 0, \tag{2.14b}$$

where $v \in X_1$ and $w \in X_2$. In cases in which the linearized operator admits certain estimates, the first equation can be solved via the implicit function theorem. In the two dimensional problems, and in the three dimensional gravity capillary wave problem, this can be done. In these cases the second equation turns out to be finite dimensional in character and it will be resolved through a reduced variational problem.

In keeping with the program outlined above we begin by identifying the relevant Banach spaces, X, C, and Y. We recall the following $L^2(T(\Gamma))$ based Sobolev spaces:

$$H^s = \{f \in L^2(T(\Gamma)) \mid \|f\|_{H^s} < \infty\}\ , \tag{2.15a}$$

where

$$\|f\|_{H^s}^2 = \sum_{k \in \Gamma'} \langle k \rangle^{2s} \left| \hat{f}(k) \right|^2\ , \tag{2.15b}$$

and

$$\langle k \rangle^2 = 1 + |k|^2\,. \tag{2.15c}$$

We also introduce the spaces $H_0^s = \{f \in H^s \mid \hat{f}(0) = 0\}$. With these in mind we prove the following lemma.

LEMMA 2.2. *Suppose that $s > \frac{n-1}{2}$. If $\sigma = 0$, then $F : H^{s+1} \times H^{s+1} \to H^s \times H_0^s$ is an analytic transformation. If $\sigma > 0$, then $F : H^{s+2} \times H^{s+1} \to H^s \times H_0^s$ is an analytic transformation.*

Proof. The proof is straightforward and relies heavily upon two facts: first, the Sobolev inequality for $s > \frac{n-1}{2}$,

$$\|uv\|_{H^s} \leq C \|u\|_{H^s} \|v\|_{H^s}\,; \tag{2.16}$$

second is the fact that the Dirichlet–Neumann operator is a bounded linear function as a function of ξ from $H^{s+1} \to H^s$, and an analytic function of η for $\eta, \xi \in H^{s+1}$. This assertion will be proven in section 5. We note that in the case $\sigma > 0$ there is a

second order derivative acting on η, while in the case $\sigma = 0$ there are only first order derivatives acting on η; this accounts for the two different spaces in the statement of the theorem. The fact that the target space of the second component of F is H^s_0 rather than H^s is a result of the following calculation:

$$\begin{aligned}\hat{F}_2(0) &= \int_{T(\Gamma)} c \cdot \partial_x \eta + G(\eta)\ \xi\ dx \\ &= \int_{T(\Gamma)} (\nabla \varphi|_\eta) \cdot N_\eta\ dx \\ &= \int_{T(\Gamma)} \int_{-h}^{\eta} \Delta \varphi\ dy\ dx \\ &= 0,\end{aligned}$$

which is true by the divergence theorem for $w = w(v, c)$ and the fact that φ is harmonic. Finally, analyticity is proven by the fact that the Dirichlet–Neumann operator is analytic, and that all other appearances of η and ξ are in an analytic fashion. □

We note that a corollary of this result is the following.

COROLLARY 2.3. *Suppose that $s > \frac{n-1}{2}$. If $\sigma = 0$, then $A(c) : H^{s+1} \times H^{s+1}_0 \to H^s \times H^s_0$ is bounded. If $\sigma > 0$, then $A(c) : H^{s+2} \times H^{s+1}_0 \to H^s \times H^s_0$ is bounded.*

We now set $X = H^{s+2} \times H^{s+1}_0$ for the case $\sigma > 0$, $X = H^{s+1} \times H^{s+1}_0$ for the case $\sigma = 0$, and $Y = H^s \times H^s_0$ in either case. In the case of the two dimensional problem we let $C = \mathbb{R}$, and in the case of the three dimensional problem we let $C = \mathbb{R}^2$.

2.3. The linearized operator. At this point we will analyze the problem (2.12) linearized around the trivial solution $(u, c) = (0, c)$. In particular we will restrict ourselves to the cases $n = 2$ and $n = 3$, and we identify the set of parameters $c \in \mathbb{R}$ (respectively, $c \in \mathbb{R}^2$) for which $A(c)$ has a nontrivial null space. Linearizing the system (2.12a) about $(u, c) = (0, c)$ involves the Dirichlet–Neumann operator $G(0) = G_0 = |D| \tanh(h|D|)$.

THEOREM 2.4. *The linearization of F about the trivial solution $(0, c)$ is given by*

$$A(c) = \partial_u F(0, c) = \begin{pmatrix} g - \sigma \Delta_x & -c \cdot \nabla_x \\ c \cdot \nabla_x & G_0 \end{pmatrix}. \tag{2.17}$$

Furthermore, given $k \in \Gamma'$, the null eigenvalues of $A(c)$ on X occur when $c \in \mathbb{R}^{n-1}$ is such that

$$\Delta_\sigma(c, k) = (g + \sigma |k|^2)\, |k| \tanh(h\, |k|) - (c \cdot k)^2 = 0\ . \tag{2.18}$$

Given c_0 satisfying (2.18) for some $k_0 \in \Gamma'$, the null space of $A(c_0)$ is even dimensional, and it is generically two dimensional. It is spanned by the eigenfunctions

$$v^1_k(x) = ((c_0 \cdot k) \cos(k \cdot x), (g + \sigma |k|^2) \sin(k \cdot x))^T, \tag{2.19a}$$

$$v^2_k(x) = (-(c_0 \cdot k) \sin(k \cdot x), (g + \sigma |k|^2) \cos(k \cdot x))^T \tag{2.19b}$$

for $k \in \Gamma'$ in the set of solutions of (2.18).

Proof. The proof is a straightforward calculation, but the role of the function $\Delta_\sigma(c, k)$ perhaps needs some explanation. We will always work in function spaces

smooth enough to admit Fourier expansions of our functions $\eta(x)$ and $\xi(x)$; therefore

$$A(c)\begin{pmatrix}\eta(x)\\ \xi(x)\end{pmatrix} = \sum_{k\in\Gamma'}\begin{pmatrix} g+\sigma|k|^2 & -ic\cdot k \\ ic\cdot k & |k|\tanh(h|k|)\end{pmatrix}\begin{pmatrix}\hat\eta(k)\\ \hat\xi(k)\end{pmatrix}e^{ik\cdot x}$$
$$= \sum_{k\in\Gamma'}\hat A_k(c)\begin{pmatrix}\hat\eta(k)\\ \hat\xi(k)\end{pmatrix}e^{ik\cdot x}\ .$$

The operator $A(c)$ is singular if one of the 2×2 blocks of $\hat A(c)$ is singular, which occurs precisely when the determinant $\Delta_\sigma(c,k)$ of the kth 2×2 block is zero, with null vectors determined by the null vector of the singular 2×2 blocks. The 2×2 block $\hat A_0(c)$ has different character, and for all $c\in\mathbb{R}^{n-1}$ it has the null eigenvector $(0,\hat\xi(0))$. It is precisely for this reason that we work in the space $X = H^{s+\rho}\times H_0^{s+1}$, so that the one dimensional subspace $\{(\eta,\xi) = \gamma(0,\hat\xi(0))\}$ does not contribute to the null space of $A(c)$ on X. Regarding the dimension of the null space, if $\hat A_k(c)$ has a null eigenvector, then so does $\hat A_{-k}(c)$; hence the null space of $A(c)$ is even dimensional. □

The relation (2.18) describes the dispersion relation between a wave number k and its phase velocity c_k. The set of solutions of (2.18) can now be described, starting with the two dimensional problem. For $k\in\Gamma'\backslash\{0\}\subseteq\mathbb{R}$ fixed, a phase velocity $c = c_k$ can always be defined through the relation

$$c^2k^2 = (g+\sigma k^2)k\tanh(hk); \tag{2.20}$$

in both cases $\sigma > 0$ and $\sigma = 0$. Therefore, for any specified spatial period Λ there is a phase velocity c such that $\Delta_\sigma(c,k) = 0$ for $k = 2\pi/\Lambda$, giving a solution to the linearized system with period Λ. Of course both $\pm k$ correspond to the same phase velocity c. It is also possible that for a given $c = c_{k_0}$ there is another $k_1\in\Gamma'$ which satisfies (2.20). Physically this corresponds to the situation where a linear gravity wave of lower wave number has the same phase velocity c as one of higher wave number. This situation can occur only if the Bond number satisfies $\sigma/gh^2 < 1/3$, where the relation (2.20) has two roots for choices of c such that $\min_k\{(g+\sigma k^2)k\tanh(hk)\} < c^2 < gh$. These two roots both satisfy $0\le k < \sqrt{g/\sigma}$. The phase velocity c_{k_0} is chosen so that one of these roots k_0 lies in the lattice Γ'. When both roots are in Γ', there is another linear solution with the specified spatial period, and the null space of $A(c_{k_0})$ is four dimensional.

The three dimensional case is only a little more complex. Given a dual lattice $\Gamma'\in\mathbb{R}^2$ and any two generators k_1, k_2 of Γ', there is always a phase velocity $c_0\in\mathbb{R}^2$ which will satisfy relation (2.18) for both. Indeed, given k_1, by the discussion of the previous paragraph we may always choose some $c_1\in\mathbb{R}^2$ which is parallel to k_1 such that (2.18) holds. Of course any other c such that $c_1\cdot k_1 = c\cdot k_1$ will also do, giving a line of solutions of (2.18) through c_1 perpendicular to k_1. The same holds for k_2, and these two lines cannot be parallel; their meeting point is the common solution that we seek. Given this common phase velocity c_0, it may be that relation (2.18) is satisfied by other wave numbers $k\in\Gamma'$ as well as for k_1 and k_2, although this situation is not generic. Indeed the relation (2.18) defines a curve in $\mathbb{R}^2$ which intersects k_1 and k_2 and possibly other points $k_3, k_4, \ldots\in\Gamma'$. This curve is symmetric about the origin, so the solutions of (2.18) appear in pairs $\pm k_\ell$. We will normalize by the choice that $k_\ell\cdot c_0 > 0$. Note that by the above discussion the situation k_ℓ perpendicular to c_0 does not occur. We have shown the following.

PROPOSITION 2.5. *Given any two generators k_1, k_2 of the lattice Γ', there is a phase velocity c_0 satisfying $\Delta_\sigma(k_\ell, c_0) = 0$, $k_\ell \cdot c_0 > 0$ for $\ell = 1, 2$, and therefore* dim ker $A(c_0) \geq 4$.

A second preliminary result is as follows.

PROPOSITION 2.6. *When $\sigma > 0$ and $c_0 \in \mathbb{R}^2$ is fixed, the number of lattice points $k \in \Gamma'$ satisfying $\Delta_\sigma(k, c_0) = 0$ is finite.*

Proof. The asymptotic behavior of the dispersion relation is that $\omega(k)^2 = (g + \sigma|k|^2)|k| \tanh(h|k|) \sim \sigma|k|^3$, while $|c \cdot k|^2 \leq C|k|^2$, and therefore the set of root of (2.18) is bounded. Notice that this finiteness result no longer holds when $\sigma = 0$. □

This paper does not address the three dimensional case with $\sigma = 0$, for the reason that it is a situation in which the phenomenon of small divisors is present. Our analysis depends upon the boundedness properties of the operator $A(c)^{-1}$ between appropriate Banach spaces. When $n \geq 3$ and $\sigma = 0$ there is a dense set of values of $c \in \mathbb{R}^{n-1}$ for which $A(c)$ has a nontrivial null space, which can even be infinite dimensional. In general, the invertibility properties of the inverse operator, even projected orthogonally to its null space, change sensitively and discontinuously with respect to c. It may be that this problem can be solved with methods related to the Nash–Moser implicit function theorem; however, there is at present no published complete proof. Among other things, these are elements of the program outlined in [21]. To describe the phenomenon of small divisors quantitatively, we give the following result.

THEOREM 2.7. *The point spectrum of the operator $A(c)$ consists of the set*

$$\left\{\mu_\pm(k) = \tfrac{1}{2}(g + |k| \tanh(h|k|)) \pm \tfrac{1}{2}\sqrt{(g - |k| \tanh(h|k|))^2 + 4(c \cdot k)^2}\right\}_{k \in \Gamma'}. \tag{2.21}$$

For every choice of the parameters g, h, and $c \in \mathbb{R}^2 \backslash \{0\}$ this set accumulates at $\mu = 0$.

Proof. Under the Fourier transform the operator $A(c)$ is 2×2 block diagonal, with eigenvalues those of the individual blocks. With $\sigma = 0$,

$$\det(\hat{A}(c)_k - \mu I) = \mu^2 - (g + |k| \tanh(h|k|))\mu + (g|k| \tanh(h|k|) - (c \cdot k)^2) \tag{2.22}$$

and the roots of this relation are the eigenvalues given above. Small divisors will occur when $\Delta_0(c, k) = \det \hat{A}(c)_k, k \in \Gamma'$ satisfies $|\Delta_0(c, k)| < |k|^{1/2}$. In this case we call $k \in \Gamma'$ a singular site in the dual lattice. The associated small eigenvalue is

$$\begin{aligned}
\mu_-(k) &= \tfrac{1}{2}(g + |k| \tanh(h|k|)) - \tfrac{1}{2}\sqrt{(g - |k| \tanh(h|k|))^2 + 4(c \cdot k)^2} \\
&= \tfrac{1}{2}(g + |k| \tanh(h|k|)) - \tfrac{1}{2}\sqrt{(g + |k| \tanh(h|k|))^2 - 4\Delta_0(c, k)} \\
&\sim \frac{\Delta_0(c, k)}{(g + |k| \tanh(h|k|))},
\end{aligned}$$

and the latter quantity is $O(|k|^{-1/2})$ for large and singular $k \in \Gamma'$.

It remains to be shown that there exist singular sites in Γ' with arbitrarily large norm. For this it suffices to consider the problem for deep water ($h = +\infty$), since $g|k| \tanh(h|k|) - g|k| \sim O(|k|e^{-2h|k|})$, and thus large singular sites for h infinite are effectively ones for finite h. For any $c \in \mathbb{R}^2 \backslash \{0\}$ choose a sequence $n_j \in \Gamma'$ which obeys an estimate $|c \cdot n_j| < d_0$; this is possible in any lattice. Additionally choose $m = m(n_j) \in \Gamma'$ such that $|c \cdot m| \sim O(|c||m|)$, $|m| < d_0|n_j|^{1/2}$, and $|m \cdot n_j| < O(|n_j|)$.

With this

$$g|n_j+m| = g(|n_j|^2+2n_j\cdot m+|m|^2)^{1/2} = g|n_j|\left(1+O\left(\frac{1}{|n_j|}\right)\right)$$

and

$$(c\cdot(n_j+m))^2 = (c\cdot m)^2+O(|m|)\ . \tag{2.23}$$

For these choices $|n_j+m| = O(|n_j|)$,

$$g|n_j+m|-(c\cdot(n_j+m))^2 = g|n_j|-((c\cdot m)^2+O(|m|))\ , \tag{2.24}$$

and by further adjustment of $m\in\Gamma'$ by choosing it such that $g|n_j+m'|-(c\cdot(n_j+m'))^2$ changes sign for some m' adjacent to m (this lies within the above constraints) we can ensure that $|g|n_j|-(c\cdot m)^2|<O(|m|)$ itself. This choice gives rise to a sequence of lattice sites $k_j = n_j+m(n_j)$ such that $|k_j|\to\infty$ and $\mu_-(k_j)\sim O(|k_j|^{-1/2})$, which is more than enough to prove the statements of the theorem. ∎

We have now identified our function spaces, the types of solutions which we seek, and the possible values of c where such bifurcation branches can be found. We must now solve the two bifurcation equations and we begin with the P equation.

3. Existence of solutions to the *P* equation. In this section we use the implicit function theorem on Banach spaces to solve the P equation of the method of Lyapunov–Schmidt. We have already established that our map $F: X\times C\to Y$ is analytic and that the linearized operator $A(c)$ is bounded from X into Y. It is clear that the projection P, being a bounded operator with finite dimensional null space, will not change either of these properties and so the only condition yet to be satisfied in order to use the implicit function theorem is the boundedness of the inverse operator. In the same way that we consider the range of $A(c)$ by operating by P, we also consider X_2 as our domain by viewing X_1 as part of the parameter space with C. This results in a solution

$$w = w(v,c), \tag{3.1}$$

such that $PF(v+w(v,c),c)=0$. With this in mind it remains to prove the following theorem.

THEOREM 3.1. *$(P\partial_w F(0,c))^{-1}$ is a bounded linear operator from Y_1 to X_2.*

Proof. We begin by identifying a likely candidate to be the inverse operator. If we considered a value of c such that $\Delta_\sigma(c,k)\neq 0$ for all $k\in\Gamma'$, then an inverse can be found by inverting the 2×2 blocks on the Fourier side resulting in

$$\hat{A}(c)_k^{-1} = \frac{1}{\Delta_\sigma(c,k)}\begin{pmatrix} |k|\tanh(h|k|) & ic\cdot k \\ -ic\cdot k & g+\sigma|k|^2\end{pmatrix}. \tag{3.2}$$

This is well defined for all $k\in\Gamma'\backslash\{0\}$. At $k=0$, $\hat{A}(c)_k$ is singular, but since the second component of F has $k=0$ mode equal to zero we can define an inverse in the following way:

$$\hat{A}(c)_0^{-1}\begin{pmatrix}\hat{v}(0)\\ 0\end{pmatrix} \equiv \begin{pmatrix}\frac{\hat{v}(0)}{g}\\ 0\end{pmatrix}. \tag{3.3}$$

We break up the rest of the proof into three parts, corresponding to each of the three cases which we are considering, since each requires a slightly different analysis.

We begin by studying the case of two dimensions, $\sigma = 0$, and a two dimensional null space. In this case $X = H^{s+1} \times H^{s+1}$. Assume that $\hat{A}(c_0)_{\pm k_0}$ is singular. By the discussion of the previous section, all other $\hat{A}(c)_k, k \neq \pm k_0$ are nonsingular. Let $\tilde{\Gamma}' = \Gamma' \backslash \{\pm k_0\}$, $\tilde{\Gamma}'_0 = \Gamma' \backslash \{0, \pm k_0\}$, and consider the estimate

$$\begin{aligned}
&\left\| (P\partial_w F(0,c))^{-1} \begin{pmatrix} u^1 \\ u^2 \end{pmatrix} \right\|_{X_2}^2 \\
&= \left| \frac{\hat{u}^1(0)}{g} \right|^2 + \sum_{k \in \tilde{\Gamma}'_0} \langle k \rangle^{2(s+1)} \left| \frac{|k| \tanh(h|k|) \hat{u}^1(k) + i c_0 k \hat{u}^2(k)}{\Delta_0(c_0, k)} \right|^2 \\
&\quad + \sum_{k \in \tilde{\Gamma}'_0} \langle k \rangle^{2(s+1)} \left| \frac{-i c_0 k \hat{u}^1(k) + g \hat{u}^2(k)}{\Delta_0(c_0, k)} \right|^2 \\
&\leq \left| \frac{\hat{u}^1(0)}{g} \right|^2 + 2 \sum_{k \in \tilde{\Gamma}'_0} \langle k \rangle^{2(s+1)} \frac{(|k| \tanh(h|k|))^2 \left| \hat{u}^1(k) \right|^2 + (c_0 k)^2 \left| \hat{u}^2(k) \right|^2}{\Delta_0(c_0, k)^2} \\
&\quad + 2 \sum_{k \in \tilde{\Gamma}'_0} \langle k \rangle^{2(s+1)} \frac{(c_0 k)^2 \left| \hat{u}^1(k) \right|^2 + g^2 \left| \hat{u}^2(k) \right|^2}{\Delta_0(c_0, k)^2} \\
&= \left| \frac{\hat{u}^1(0)}{g} \right|^2 + 2 \sum_{k \in \tilde{\Gamma}'_0} \langle k \rangle^{2s} \frac{\langle k \rangle^2 \left[(|k| \tanh(h|k|))^2 \left| \hat{u}^1(k) \right|^2 + (c_0 k)^2 \left| \hat{u}^2(k) \right|^2 \right]}{\Delta_0(c_0, k)^2} \\
&\quad + 2 \sum_{k \in \tilde{\Gamma}'_0} \langle k \rangle^{2s} \frac{\langle k \rangle^2 \left[(c_0 k)^2 \left| \hat{u}^1(k) \right|^2 + g^2 \left| \hat{u}^2(k) \right|^2 \right]}{\Delta_0(c_0, k)^2}.
\end{aligned}$$

Now, by using the fact that we can bound $\Delta_0(c_0, k)$ from below by $K_1 + K_2 \langle k \rangle^2$ we continue the estimate,

$$\begin{aligned}
&\left\| (P\partial_w F(0,c))^{-1} \begin{pmatrix} u^1 \\ u^2 \end{pmatrix} \right\|_{X_2}^2 \\
&\leq \left| \frac{\hat{u}^1(0)}{g} \right|^2 + C_1 \sum_{k \in \tilde{\Gamma}'_0} \langle k \rangle^{2s} \left| \hat{u}^1(k) \right|^2 + C_2 \sum_{k \in \tilde{\Gamma}'_0} \langle k \rangle^{2s} \left| \hat{u}^2(k) \right|^2 \\
&\leq C \left[\sum_{k \in \tilde{\Gamma}'} \langle k \rangle^{2s} \left| \hat{u}^1(k) \right|^2 + \sum_{k \in \tilde{\Gamma}'_0} \langle k \rangle^{2s} \left| \hat{u}^2(k) \right|^2 \right] \\
&= C \left\| \begin{pmatrix} u^1 \\ u^2 \end{pmatrix} \right\|_{Y_1}^2.
\end{aligned}$$

The case of $\sigma > 0$ in two dimensions is very similar, and proceeding in the same way as before we can produce an estimate of the form

$$\left\| (P\partial_{u_2} F(0,c))^{-1} \begin{pmatrix} u^1 \\ u^2 \end{pmatrix} \right\|_{X_2}^2$$

$$\leq \left| \frac{\hat{u}^1(0)}{g} \right|^2 + 2 \sum_{k \in \tilde{\Gamma}'_0} \langle k \rangle^{2s} \frac{\langle k \rangle^4 \left[(|k| \tanh(h|k|))^2 \left| \hat{u}^1(k) \right|^2 + (c_0 k)^2 \left| \hat{u}^2(k) \right|^2 \right]}{\Delta_\sigma(c_0, k)^2}$$

$$+ 2 \sum_{k \in \tilde{\Gamma}'_0} \langle k \rangle^{2s} \frac{\langle k \rangle^2 \left[(c_0 k)^2 \left| \hat{u}^1(k) \right|^2 + (g + \sigma |k|^2)^2 \left| \hat{u}^2(k) \right|^2 \right]}{\Delta_\sigma(c_0, k)^2}.$$

Now using the fact that we can produce constants K_3 and K_4 such that $|\Delta_\sigma| \geq K_3 + K_4 \langle k \rangle^3$ we again obtain the estimate desired. The proofs in the cases of higher dimensional null spaces and three dimensions ($\sigma > 0$) are almost the same as the previous except that one has to account for a possibly larger null space for the linearized operator $A(c)$. The sums are over $\tilde{\Gamma}'_0 = \Gamma' \backslash \{0, \pm k_j\}_{j=1}^N$, which serves to omit from them any of the wave numbers k_j such that $\Delta_\sigma(c, k_j) = 0$. Again, the constants K_3 and K_4 can be produced due to the fact that for $|k|$ large enough, $\Delta_\sigma(c, k) > 0$. □

With these estimates in hand it is easy to prove the following theorem.

THEOREM 3.2. *The equation $PF(v + w, c) = 0$ has a solution*

$$w = w(v, c) \tag{3.4}$$

for all (v, c) in a ball $B_\varepsilon(0, c_0)$ which is locally unique, such that

$$PF(v + w(v, c), c) = 0, \tag{3.5a}$$

$$w(0, c) = 0. \tag{3.5b}$$

Furthermore, w is analytic as a function of (v, c), $\|w(v, c)\|_X \leq C \|v\|_X^2$, and it is equivariant under the torus action $T_\alpha w(v, c) = w(T_\alpha v, c)$.

Proof. We use the implicit function theorem in conjunction with Theorem 3.1. □

4. The reduced variational problem. With the choice of $u = v + w(v, c)$ (2.14a) is solved, and we focus on solutions (v, c) which will also result in a solution of (2.14b). The approach that we use is close to that of J. Moser [17], with the novelty that the problem is equivariant with respect to the action of a torus $\mathbb{T}^2$. The variational problem that we solve is equivalent to finding critical points of the Hamiltonian function $H(u)$ of (2.8), when restricted to the subset of phase space $\{u \in X \ : \ I_1(u) = a_1, I_2(u) = a_2\}$. This is not feasible by direct variational methods. Instead, following [17], we pose a reduced variational problem in the finite dimensional space X_1, whose solutions will solve (2.14b). In preparation for this we make an astute choice of $c = c(v)$. The function $u = v + w(v, c)$ solves (2.14a); therefore

$$(\delta H - c \cdot \delta I)(v + w(v, c)) = q(v, c) \in Y_2 \ . \tag{4.1}$$

Let $\{\pm k_1, \ldots \pm k_N\} = K \subseteq \Gamma'$ be the collection of wave numbers for which $\Delta(c_0, k_p) = 0$, which are normalized so that $c_0 \cdot k_p > 0$. We also choose the norm $|v|$ so that it is invariant under the torus action $v \to T_\alpha v$.

LEMMA 4.1. *Suppose that the vector $a = (a_1, a_2)$ is not collinear with any $k_p \in K$, and that $|(a_1, a_2)| < \delta$ for sufficiently small δ. On the set $\{(v, c) \ : \ I_1(v + w(v, c)) = a_1, I_2(v + w(v, c)) = a_2\} \subseteq X_1 \times \mathbb{R}^2$ we can make a choice of $c = c(v)$ to satisfy*

$$\int_{T(\Gamma)} (q(v, c) \ , \ \delta_u I_j(v + w(v, c))) \ dx = 0 \ , \qquad j = 1, 2 \ . \tag{4.2}$$

Furthermore, $c(v)$ is invariant under the action T_α of the torus, $c(T_\alpha v) = c(v)$; $|c(v) - c_0| = O(|v|^2)$; and for $v \neq 0$ $c(v)$ is real analytic in v.

We will defer the proof of this for several pages, for the sake of continuity of the argument. Using both $c(v)$ and $w(v, c(v))$, define the functionals

$$I_j^Q(v) = I_j(v + w(v, c(v))) \ , \qquad j = 1, 2 \ , \tag{4.3}$$

and

$$H^Q(v) = H(v + w(v, c(v))) \ . \tag{4.4}$$

These functionals are invariant under the torus action $T_\alpha : X_1 \to X_1$. For $a = (a_1, a_2)$ not collinear with any of the set K, and for $|a| < \delta$ sufficiently small, consider the subset of the space X_1

$$S(a) = \{v \in X_1 \ : \ I_1^Q(v) = a_1 \ , \ I_2^Q(v) = a_2\} \ . \tag{4.5}$$

Clearly $S(a)$ is also invariant under T_α, and it is a smooth submanifold of X_1, to which Lemma 4.1 applies. Finally define the reduced action functional

$$A(v) = H^Q(v) - c(v) \cdot (I^Q(v) - a) \ . \tag{4.6}$$

THEOREM 4.2. *When $\delta_v I_1^Q$ and $\delta_v I_2^Q$ are linearily independent at all $v \in S(a)$, then $S(a)$ is a submanifold of X_1 of codimension 2. In this case, critical points of $A(v)$ on $S(a)$ correspond to solutions $u = v + w(v, c(v))$, $c(v)$ of the bifurcation equation $QF = 0$ of* (2.14b).

Proof. Choices of (v, c) such that $q(v, c) = 0$ correspond precisely to solutions of the desired equation (2.14b). The choice of $c(v)$ prescribed in Lemma 4.1 is so that q is orthogonal in X_1 to the subspace of normal vectors $N_v(S(a))$. Indeed with $q \in X_1$, and a basis of $N_v(S(a))$ given by $\{\delta_v I_1^Q, \delta_v I_2^Q\}$,

$$\begin{aligned} \int_{T(\Gamma)} (q, \delta_v I_j^Q) dx &= \int_{T(\Gamma)} (q, Q \delta_u I_j(v + w(v, c))) dx \\ &= \int_{T(\Gamma)} (q, \delta_u I_j) dx \\ &= 0 \ . \end{aligned}$$

Now suppose that v is a critical point on $S(a)$ of the functional $A(v)$. For any $\delta v \in T_v(S(a))$ the tangent space, we have the calculation

$$\begin{aligned} \partial_v A(v) \cdot \delta v = &\int_{T(\Gamma)} Q(\delta_u H(v + w) - c(v) \cdot \delta_u I(v + w)) \delta v \ dx \\ &+ \int_{T(\Gamma)} P(\delta_u H(v + w) - c(v) \cdot \delta_u I(v + w))(\partial_v w \cdot \delta v) \ dx \\ = &\ 0, \end{aligned}$$

where the term containing $\partial_v c(v)$ has dropped out as the expression is evaluated on $S(a)$. Furthermore $(\partial_v w \cdot \delta v) \in X_2$, and we have resolved the first bifurcation equation (2.14a), so that the second term is zero as well. Thus for $v \in S(a)$ and $\delta v \in T_v(S(a))$, $\partial_v A(v) \cdot \delta v = \int_{T(\Gamma)} (q, \delta v) dx$, and at critical points it vanishes. Since $c = c(v)$ is chosen so that $\int_{T(\Gamma)} (q, \delta v) dx = 0$ for all $\delta v \perp T_v(S(a))$ as well, indeed critical points

correspond to zeros of $q = q(v, c(v))$. Because $A(v)$ is invariant under the action T_α, so is the set of critical points on $S(a)$. □

At this point in the argument, any choices of a such that $S(a)$ is compact will yield at least one solution of the Euler equations (2.9) ($S(a)$ is a point when $m = 2$). We will, however, be able to use the topology of the quotient space $S(a)/\mathbb{T}^2$ to obtain a more precise and generally larger lower bound on the number of critical orbits of $A(v)$ on $S(a)$.

4.1. Application of a T^2 equivariant cohomological index. Parameterize $v \in X_1$ by $\mathbb{C}^N$ by setting $z = (r_1 e^{i\varphi_1}, \ldots, r_N e^{i\varphi_N})$, and

$$v = \sum_{p=1}^{N} r_p(\cos(\varphi_p) v^1_{k_p} + \sin(\varphi_p) v^2_{k_p}) \ . \tag{4.7}$$

The torus action T_α on X_1 is defined by

$$T_\alpha v = \sum_{p=1}^{N} r_p(\cos(\varphi_p + k_p \cdot \alpha) v^1_{k_p} + \sin(\varphi_p + k_p \cdot \alpha) v^2_{k_p}), \tag{4.8}$$

which in $\mathbb{C}^N$ is efficiently described as $T_\alpha z = (r_p e^{(i\varphi_p + k_p \cdot \alpha)})_{p=1}^N$. A calculation yields that

$$I^Q_j(v) = \frac{1}{2} \int_{T(\Gamma)} (v, J\partial_{x_j} v)\, dx = \sum_{p=1}^{N} k^j_p r^2_p (c_0 \cdot k_p (g + \sigma |k_p|^2)) |T(\Gamma)| \tag{4.9}$$

for $j = 1, 2$ where $J = \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$. Note that we have previously normalized wave numbers so that $c_0 \cdot k_p > 0$, $1 \le p \le N$. We will use the notation $c(p) = c_0 \cdot k_p (g + \sigma |k_p|^2) |T(\Gamma)|$.

Let us start the discussion with the nonresonant case $N = 2$, with $k_1, k_2 \in \Gamma'$ not collinear. Choose a 2×2 change of basis M so that $Mk_1 = (1, 0)^T$ and $Mk_2 = (0, 1)^T$. This gives an equivalent definition of $S(a) = \{v \in X_1 \ : \ MI = Ma = b\}$ and modifies the torus action to a product action $T_\beta z = (z_1 e^{i\beta_1}, z_2 e^{i\beta_2})$, where $M^T \beta = \alpha$. Clearly $S(a)$ is the two dimensional torus

$$S(a) = \{v \ : \ r^2_1 c'(1) = b_1 \ , \ r^2_2 c'(1) = b_2\} \tag{4.10}$$

for positive constants $c'(p) = \sum_q M_{pq} c(q)$, and the orbit under T_β of any point of $S(a)$ consists of the whole set. Therefore the Hamiltonian H is constant on the set, and we have proved the following result.

THEOREM 4.3. *In case $N = 2$ and k_1, k_2 not collinear, whenever a is not collinear with either k_1 or k_2, and $|a| < \delta$, each submanifold $S(a) \in X_1$ corresponds to a solution of* (2.9) *and its translates by T_α. Furthermore the family of solutions $u = v + w(v, c(v)) \in X$, $v \neq 0$, is locally real analytic, forming a four dimensional bifurcation manifold which is invariant under the action of T_α.*

The theorem is the analogue of the original Lyapunov center theorem, in which there are no allowances for resonances. The statement of analyticity is an immediate consequence of the regularity results of Theorem 3.2 and Lemma 4.1, stemming ultimately from the implicit function theorem. In the particular case that $k_1 = (k^1_1, k^2_1)$, and $k_2 = (k^1_1, -k^2_1)$, with $k^1_1 \neq k^2_1$, the curve $\{a_1 = a_2\} \in S(a)$ consists of the symmetric diamond solutions of the three dimensional water wave problem in J. Reeder and M. Shinbrot [23].

Now turn to the resonant case $N \geq 2$. We will consider two possibilities separately, either (i) $N > 2$, for which necessarily there are two noncollinear wave vectors among $\{k_p\}_{p=1}^N$, or else (ii) $N = 2$ and k_1 and k_2 are linearly dependent. The latter case corresponds to the two dimensional resonant problem, with resonant interaction between periodic gravity waves and capillary waves with the same phase velocity; this was previously discussed in M. Jones and J. Toland's work [13], [14]. In case (ii) we will give an independent proof of the existence of at least two distinct periodic solutions; the proof will be put off until the end of the present section.

In case (i) of the preceding paragraph, in which $N > 2$, the wave vectors $\{k_p\}_{p=1}^N$ involved in the kernel X_1 lie within the cone $\{c_0 \cdot k > 0\}$. Choose k_1 to be the leftmost of this collection, k_N to be the rightmost. In case either or both of these are collinear with another from the collection (from our discussion of the dispersion relation in section 2.3, at most two wave vectors can be collinear), take k_1 and/or k_N to be the smaller, and k_2 and/or k_{N-1} to be the bigger. As in the nonresonant case we may make a change of basis M so that $k_1 = (1,0)^T$ and $k_N = (0,1)^T$; without changing notation assume that this is so. Then all other wave vectors are expressed as $k_p = k_p^1 k_1 + k_p^2 k_N$ for rationals $k_p^\ell > 0$, except that in case of collinearity it may be that $k_2 = (k_2^1, 0)^T$ and/or $k_{N-1} = (0, k_{N-1}^2)^T$. All of the other coefficients are positive from this exercise, as k_p lie in the positive cone of the new basis $\{k_1, k_N\}$. We also order the remaining wave vectors $k_p = (k_p^1, k_p^2)^T$ in terms of increasing slope; $k_p^2/k_p^1 \leq k_{p+1}^2/k_{p+1}^1$. It again follows from section 2.3 that at most two wave vectors can have a common slope.

THEOREM 4.4. *Suppose that for all $1 \leq p \leq N$, $k_p^2/k_p^1 \neq a_2/a_1$, and that $|a| < \varepsilon^2$ for sufficiently small ε. Then the set $S(a) = \{v \in X_1 \ : \ I_1 = a_1, \ I_2 = a_2\}$ is a compact submanifold of X_1.*

Proof. From the form (4.9) of the action integrals I_j^Q, it is clear that the gradients $\delta_v I_1^Q$ and $\delta_v I_2^Q$ are independent except when (a_1, a_2) is taken so that two semiaxes of the cylindrical ellipsoids $\{I_1^Q = a_1\}$ and $\{I_2^Q = a_2\}$ coincide, and we can invoke Theorem 4.2. There is in fact more structure than this. When $a_2/a_1 = k_p^2/k_p^1$, and k_p is the sole wave vector with this slope, the intersection must be at $\{z_j = 0, \ j \neq p\} \cap S(a)$, which is a circle (invariant under the action of T_α). In case of multiplicity two, this is at $\{z_j = 0, \ j \neq p, p+1\} \cap S(a)$, which is a T_α invariant ellipsoid homeomorphic to S^3. □

With no further work one can see that for $N > 2$ and for choices of a avoiding the singular cases $\{k_p^2/k_p^1 \ : \ p = 1, \ldots, N\}$ there are at least two distinct solutions of (2.9), corresponding to the maximum and minimum of the reduced Hamiltonian H^Q on $S(a)$. In all cases in which both $a_1, a_2 > 0$ these are truly three dimensional solutions, as they have a nonzero component of $v_p \in X_1$ for at least two linearly independent wave vectors k_p. Indeed, if k_p is the only wave number having slope k_p^2/k_p^1, and if $a_2/a_1 \neq k_p^2/k_p^1$, then $\{z_j = 0 \ : j \neq p\} \cap S(a)$ is empty. In case k_p is collinear with k_{p+1}, but still $a_2/a_1 \neq k_p^2/k_p^1$, then as well $\{z_j = 0 \ : j \neq p, p+1\} \cap S(a) = \emptyset$, since if $z = (0, \ldots z_p, z_{p+1}, \ldots, 0) \in S(a)$, then

$$|z_p|^2 + \frac{k_{p+1}^1}{k_p^1}|z_{p+1}|^2 = \frac{a_1}{k_p^1} = |z_p|^2 + \frac{k_{p+1}^2}{k_p^2}|z_{p+1}|^2 = \frac{a_2}{k_p^2} \ , \tag{4.11}$$

giving the contradiction $a_2/a_1 = k_p^2/k_p^1$. This argument shows that solutions of (2.9) associated with $v \in S(a)$ are genuinely three dimensional in character.

THEOREM 4.5. *Given $v \in S(a)$ with $a_2/a_1 \notin \{k_p^2/k_p^1 \ : \ p = 1, \ldots, N\}$, then v contains nonzero Fourier modes for at least two wave vectors which are not collinear.*

There is further topology of the orbit space $S(a)/T^2$ which will, in cases of resonance $N > 2$, guarantee, in general, that there are more solutions. This will be shown by an argument involving a cohomological index equivariant with respect to the group action of a torus, similar to the topological category argument of J. Moser [17] with the action of a circle. In what follows, $\mathrm{ind}_{T_\alpha}(S)$ will be a T_α equivariant cohomological index of the set S. The following theorem gives a lower bound for distinct critical orbits of a T_α invariant function H^Q on $S(a)$ in this situation, in terms of the index. It furthermore provides an estimate of the index, which is sharp.

THEOREM 4.6. *Let $N > 2$ and choose a as in* Theorem 4.4 *so that $S(a)$ is a compact manifold. Then there exist at least* $\mathrm{ind}_{T_\alpha}(S(a)) + 1$ *distinct critical orbits of H^Q on $S(a)$, corresponding to distinct solutions of* (2.9). *We can furthermore make a choice of the index so that* $\mathrm{ind}_{T_\alpha}(S(a)) = N - 2$.

A standard argument implies that the number of critical points of invariant functionals H on the set $S(a)$ is bounded below by $\mathrm{ind}_{T_\alpha}(S(a)) + 1$. The crux of the proof is to show that, while the topology of $S(a)$ and $S(a)/T^2$ varies depending upon a, in all nonsingular cases one can define an index so that $\mathrm{ind}_{T_\alpha}(S(a)) = N - 2$.

We claim that for all a such that $a_2/a_1 \notin \{k_p^2/k_p^1 \ : \ p = 1, \ldots, N\}$, then $S(a) \simeq S^{2d-1} \times S^{2(N-d)-1}$. Indeed, define

$$J_1 = \frac{1}{a_1} I_1 + \frac{1}{a_2} I_2 = \sum_{p=1}^{N} c(p) \left(\frac{k_p^1}{a_1} + \frac{k_p^2}{a_2} \right) r_p^2$$

and

$$J_2 = \frac{1}{a_1} I_1 - \frac{1}{a_2} I_2 = \sum_{p=1}^{N} c(p) \left(\frac{k_p^1}{a_1} - \frac{k_p^2}{a_2} \right) r_p^2;$$

an equivalent description of $S(a)$ is that

$$S(a) = \{ v \in X_1 \ : \ J_1 = 2, \ J_2 = 0 \} \ . \tag{4.12}$$

Suppose that $k_d^2/k_d^1 < a_2/a_1 < k_{d+1}^2/k_{d+1}^1$, then for $1 \leq p \leq d$ the coefficients $(k_p^1/a_1 - k_p^2/a_2)$ of J_2 are positive, and they are negative for $d+1 \leq p \leq N$. We may write $P = \sum_{p=1}^{d} c(p)(\frac{k_p^1}{a_1} - \frac{k_p^2}{a_2}) r_p^2$, $N = -\sum_{p=d+1}^{N} c(p)(\frac{k_p^1}{a_1} - \frac{k_p^2}{a_2}) r_p^2$; then $J_2(v) = P - N$ and the set $S(a)$ is homeomorphic through radial projections to the manifold

$$\begin{aligned} \tilde{S}(a) &= \{ z \in \mathbb{C}^N \ : \ (P+N)(z) = 2 \ , \ \ (P-N)(z) = 0 \} \\ &= \{ z \in \mathbb{C}^N \ : \ P(z) = 1 \ \text{ and } \ N(z) = 1 \} \ . \end{aligned}$$

This expression makes it is clear that $\tilde{S}(a) \simeq S^{2d-1} \times S^{2(N-d)-1}$, a product of odd dimensional spheres.

We are interested in functions on $S(a)$ which are invariant under the action of T_α; therefore, it is essentially the index of the orbit space $S(a)/T^2 \simeq S^{2d-1} \times S^{2(N-d)-1}/T^2$ which is relevant. As a first example, we will suppose that T_α is a product T^2 action on the manifold $S^{2d-1} \times S^{2(N-d)-1}$, and that it is free (although in the capillary gravity wave problem this is almost never the case);

$$T_\alpha(z_1, \ldots, z_N) = (z_p e^{i\alpha_1 k^1})_{p=1}^{d} (z_p e^{i\alpha_2 k^2})_{p=d+1}^{N} \ . \tag{4.13}$$

Therefore $S(a)$ is a manifold which factors into two odd dimensional spheres, each with a circle action which is equivalent to the Hopf fibration. That is, the orbit space

is $S(a)/T^2 \simeq \mathbb{C}P(d-1) \times \mathbb{C}P(N-d-1)$. In this case we will take the equivariant cohomological index $\operatorname{ind}_{T_\alpha}(S(a))$ to coincide with the Čech cohomology cuplength of the quotient manifold; $\operatorname{cuplength}(S(a)/T^2)$. The following proposition consists of standard facts about the notion of index, Liusternik–Schnirlman category, and cuplength.

PROPOSITION 4.7. *The number of critical points of a C^2 function H on M is bounded below by* $\operatorname{Cat}(M)$. *Furthermore,*

(i) $\operatorname{Cat}(M) \geq \operatorname{cuplength}(M) + 1 \geq \operatorname{ind}_{T_\alpha}(M) + 1$;

(ii) $\operatorname{cuplength}(M_1 \times M_2) = \operatorname{cuplength}(M_1) + \operatorname{cuplength}(M_2)$;

(iii) *when* $M \simeq \mathbb{C}P(q)$, *then* $\operatorname{Cat}(M) = \operatorname{cuplength}(M) + 1 = q + 1$.

We deduce from (ii) *and* (iii) *that*

(iv) $\operatorname{Cat}(\mathbb{C}P(d-1) \times \mathbb{C}P(N-d-1)) \geq N - 1$.

The remaining task is to produce the same lower bound in the general case, in which the torus action T_α on $S(a)$ is not free, and when the action of T_α is as a twisted product. We note that, while T_α is not a free action, at least the stabilizers $s(z) = \{\alpha \in T^2 \ : \ T_\alpha z = z\}$ are always finite. The following argument is an adaptation of the classical one of A. Borel, which we learned from T. Goodwillie and D. Sinha.

The first step in the construction of a T_α equivariant cohomology for $S(a)$ is to identify a universal total space $E = S^\infty \times S^\infty$ and the classifying space $E/T^2 = B_{T^2} \sim \mathbb{C}P^\infty \times \mathbb{C}P^\infty$; see, for example, E. Fadell and P. Rabinowitz [9] or Rabinowitz [22]. The action of T_α on $E \times S(a)$ is free, and because E is contractible and the only isotropy subgroups which appear are finite, the space $S(a)/T^2$ and the homotopy orbit space $(E \times S(a))/T^2$ have the same rational cohomology.

The cohomology of the classifying space B_{T^2} is generated by two elements $y_1 = y \otimes 1$ and $y_2 = 1 \otimes y$, where y is the generator of the cohomology of $\mathbb{C}P^\infty$. Because of the presence of more than one generator, the construction of [9] will not work directly. We note, however, that in the same vein a well-defined index based on cuplength can be defined for any graded subalgebra of $H^*(B;\mathbb{Q})$. Let Y be a finitely generated graded subalgebra of $H^*(B;\mathbb{Q})$ with a basis $\{y_1, \ldots, y_q\}$, and let $F \ : \ (E \times S(a))/T^2 \to B_{T^2}$ be a classifying map. We define an equivariant cohomological index to be

$$\operatorname{ind}_Y((E \times S(a))/T^2) = \max\{|k| \ : \ k \in \mathbb{N}^q, F^*(y^k) \neq 0\} \ .$$

This index coincides with that of E. Fadell and P. Rabinowitz [9] when Y is generated by one element, and with the usual cuplength when $Y = H^*(B;\mathbb{Q})$. In our situation we will take Y to be generated by $\{y_1, y_2\}$, which is the latter case.

The rational cohomology of the homotopy orbit space $(E \times S(a))/T^2$ is obtained using the Leray–Serre spectral sequences. This calculation can be paraphrased in terms of Poincaré series. The E_2 page of the spectral sequence is $E_2^{p,q} = H^p(B_{T^2}) \otimes H^q(S(a))$, whose associated Poincaré series is

$$P_B(t) = (1 - t^2)^{-2} \ .$$

One nontrivial differential occurs at the E_{2p} page, and another at the $E_{2(N-p)}$ page (or in the inverse order, if $N/2 < p$), resulting in the Poincaré polynomial for the quotient

$$\begin{aligned} P_{(E\times S(a))/T^2}(t) &= \frac{(1-t^{2p})(1-t^{2(N-p)})}{(1-t^2)^2} \\ &= (1 + t^2 + \cdots + t^{2(p-1)})(1 + t^2 + \cdots + t^{2(N-p-1)}) \ . \end{aligned}$$

The coefficients of $P_{(E\times S(a))/T^2}(t)$ give the Betti numbers β_j of $S(a)/T^2$, as its rational cohomology coincides with that of the homotopy orbit space. From this expression, $\beta_{2(N-2)} = 1$; hence $H^{2(N-2)}((E \times S(a))/T^2; \mathbb{Q})$ is one dimensional, and there is some nonzero element in the coset of $\{y_1^\gamma y_2^{N-2-\gamma}\}_{\gamma=0}^{N-2}$. All higher cohomology classes vanish. We deduce that there are classes $x_1, x_2 \in H^2$ for which the cup product $x_1^\gamma x_2^{N-2-\gamma}$ is nonzero, so that $\operatorname{ind}_{T_\alpha}((E \times S(a))/T^2) = N - 2$. Applying Proposition 4.7, we conclude the result of Theorem 4.6.

We note for further interest that for Morse functions H on $S(a)/T^2$, the Morse inequalities state that the number of critical points is at least $P_{(E\times S(a))/T^2}(1) = \sum_{j=1}^{2(N-2)} \beta_j = p(N-p)$. This estimate exceeds $N-1$, strictly so if $p \neq 1, N-1$. The equivariant cohomology of products of spheres under torus actions is relevant in other problems, in particular in the problem of resonant tori in dynamical systems; it is discussed further in W. Craig and D. Haskell [6].

4.2. Two dimensional capillary gravity waves. The point of this short subsection is to provide an alternate proof of the existence of gravity waves in two dimensions [15], [27], and capillary gravity waves in two dimensions, both in the regular case [31], [2], [1] and in the presence of resonance [13], [14]. This can be done in the present framework, considering the circle $T(\Gamma) = \mathbb{R}/\Gamma$ and setting our basic function space to be $X = H^s(T(\Gamma))$. The null space of the linearized operator A is denoted by X_1; in the nonresonant case it is two dimensional, and in the case of resonance it is four dimensional. By Theorem 3.2 we obtain a solution to (2.14a) in a ball $B_\mathcal{E}(0, c_0)$ which is analytic in both variables (v, c). The relevant functionals for the variational problem are

$$I_1 = \int_{S^1} \eta \partial_{x_1} \xi \, dx \ , \tag{4.14}$$

and H gives as in (2.8) by a one dimensional integral. In the two dimensional problem the group action is a circle action on $S(a)$ which leaves the functions H, I invariant. Furthermore, the one dimensional version of Lemma 4.1 holds [16, Lemma 2, p. 739], giving an analytic function $c = c(v)$ such that $q = F(v + w(v, c(v)), c(v))$ is in the tangent space of $S(a) = \{v \in X_1 \ : \ I^Q(v) = a\}$. The object is to solve (2.14b) using the reduced variational problem.

In the nonresonant case, $N = 1$ and each subset $S(a) \in X_1, a > 0$, is a one dimensional submanifold consisting entirely of solutions of (2.9), related to each other by translation by T_α; this is the analog of Theorem 4.3. The solutions are in fact analytic in v, as in the classical Lyapunov center theorem.

We now consider the resonant case. Let us suppose that the null space of $A(c_0)$ is generated by Fourier modes $k_1, k_2 = Rk_1 \in \Gamma' \simeq \mathbb{Z}$. For small Bond number, we have in fact $k_1 \ll k_2$. The subset of X_1 with fixed momentum is

$$S(a) = \{v \in X_1 \ : \ I_1 = c(1)|z_1|^2 + c(2)|z_2|^2 = a\} \ . \tag{4.15}$$

The circle action on $S(a)$ is given by $T_\alpha z = (z_1 e^{i\alpha}, z_2 e^{iR\alpha})$, and incidentally the quotient $S(a)/T_\alpha$ is never a manifold. Nonetheless, the submanifold $S(a) \simeq S^3$ is compact, and therefore the function H is either constant, or else has at least two distinct critical values, namely, its maximum and its minimum. This gives rise to at least two distinct critical circles of H on $S(a)$, invariant under T_α.

THEOREM 4.8. *The two dimensional gravity wave problem and the two dimensional capillary gravity wave problem have analytic families of periodic solutions bifurcating from each c_k, $k \in \Gamma'$, for which the null space X_1 is two dimensional. In*

case of resonance in the capillary gravity wave problem, X_1 is four dimensional and there are at least two distinct solutions on every level surface $S(a)$ of the momentum I_1, for sufficiently small $a > 0$.

We remark that we would not obtain a better lower estimate for the number of critical orbits of H on $S(a)$ from an index theory argument, or even from a more refined argument by Morse theory, if indeed we were to show that H was a Morse function on $S(a)/T^2$.

4.3. Proof of Lemma 4.1. To finish this section we give a proof of Lemma 4.1, which is based on the implicit function theorem. From Theorem 3.2 we have the estimate $\|w(v,c)\|_X \leq C\,\|v\|_X^2$, and we also find

$$\begin{aligned}\|q\|_Y &= \|\delta_u H(v+w) - c\cdot\delta_u I(v+w)\|_Y \\ &\leq \big\|Q(\delta_u^2 H(0)(v+w) - c_0\cdot\delta_u I(v+w))\big\|_Y \\ &\quad + \|Q(\delta_u H_3(v+w) - (c-c_0)\cdot\delta_u I(v+w))\|_Y \\ &= \|Q(\delta_u H_3(v+w) - (c-c_0)\cdot\delta_u I(v))\|_Y \ .\end{aligned}$$

The notation is that $I = (I_1, I_2)$, that $\delta_u H(u) = \delta_u^2 H(0)u + \delta_u H_3(u)$ describes the Taylor series expansion with remainder, and we have used that Q projects orthogonally in X onto an invariant subspace of J (a symplectic subspace) so that $Q\delta_u I(u) = \delta_u I(Qu)$. Letting $v = \sum_{p=1}^N r_p[\cos(\varphi_p)v_{k_p}^1 + \sin(\varphi_p)v_{k_p}^2]$ as in (4.7), we compute the quantities

$$\begin{aligned}\delta_u I_\ell(v) &= \sum_{p=1}^N r_p(\cos(\varphi_p)J\partial_{x_\ell}v_{k_p}^1 + \sin(\varphi_p)J\partial_{x_\ell}v_{k_p}^2) \\ &= \sum_{p=1}^N r_p\cos(\varphi_p)J\partial_{x_\ell}\begin{pmatrix} c_0\cdot k_p\cos(k_p\cdot x) \\ (g+\sigma|k_p|^2)\sin(k_p\cdot x)\end{pmatrix} \\ &\quad + r_p\sin(\varphi_p)J\partial_{x_\ell}\begin{pmatrix} -c_0\cdot k_p\sin(k_p\cdot x) \\ (g+\sigma|k_p|^2)\cos(k_p\cdot x)\end{pmatrix}\end{aligned}$$

and

$$(4.16)\quad \int_{T(\Gamma)}(\delta_u I_\ell(v), \delta_u I_j(v))\ dx = \sum_{p=1}^N r_p^2(k_p^j k_p^\ell)|T(\Gamma)|((c_0\cdot k_p)^2 + (g+\sigma|k_p|^2)^2)\ .$$

Let us suppose that $a = (a_1, a_2)$ is not collinear with any wave vector in the set K, and that (v,c) satisfies $I(v + w(v,c)) = a$. Define the quantity

$$\begin{aligned}m(v,c) &= \frac{1}{|v|^2}\int_{T(\Gamma)}(q(v,c)\ ,\ \delta_u I_j(v+w(v,c)))\ dx \\ &= \frac{1}{|v|^2}\left(\int_{T(\Gamma)} -\sum_{\ell=1}^2 (c-c_0)_\ell(\delta_u I_\ell(v), \delta_u I_j(v))dx + O(|v|^3)\right)\ ,\end{aligned}$$

which maps $X_1\times\mathbb{R}^2$ to $\mathbb{R}^2$, and whose vanishing implies (4.2). To analyze the set $m(v,c) = 0$, evaluate m on a line $\{v = \rho e : \rho\in\mathbb{R}, |e|_{X_1} = 1\}$ through the origin, setting $m_e(\rho, c) = m(\rho e, c)$. Clearly $m_e(0, c_0) = 0$, and taking the derivative with

respect to c,

$$\partial_{c_\ell} m_{e,j}(\rho, c) = -\frac{1}{|v|^2} \int_{T(\Gamma)} (\delta_u I_\ell(v), \delta_u I_j(v)) dx|_{v=\rho e} + O(|\rho|)$$
$$= -\int_{T(\Gamma)} (\delta_u I_\ell(e), \delta_u I_j(e)) dx + O(|\rho|) \ .$$

If this Jacobian is invertible, the implicit function theorem implies the existence of a solution $c(\rho)$ of $m_e(\rho, c) = 0$ in a sufficiently small neighborhood of the origin. From (4.16) we see that $\partial_{c_\ell} m_{ej}(0, c)$ is a sum with nonnegative coefficients of symmetric rank one matrices $k_p k_p^T$. As long as two linearly independent k_p of the sum have their respective coefficients nonzero, it is invertible, and the statement of the lemma holds for sufficiently small ε. Let us suppose then that $\partial_c m_e(0, c)$ is not invertible, for purposes of contradiction. Then both

$$\partial_{c_j} m_{e,\ell}(0, c_0) = -\sum_s C_{p_s}^{(1)} k_{p_s}^j k_{p_s}^\ell \ ,$$
$$I_j(e) = \sum_s C_{p_s}^{(2)} k_{p_s}^j,$$

where all k_{p_s} are collinear, and where the coefficients are $C_{p_s}^{(1)} = ((c_0 \cdot k_p)^2 + (g + \sigma|k_p|^2)^2)|T(\Gamma)|$, $C_{p_s}^{(2)} = ((c_0 \cdot k_p)(g + \sigma|k_p|^2))|T(\Gamma)|$. However, for $v = \rho e$, we have $(a_1, a_2) = (I_1(v + w(v, c)), I_2(v + w(v, c))) = \rho^2 (I_1(e), I_2(e)) + O(|\rho|^4)$ and this in turn is $\rho^2 \sum_s C_{p_s}^{(2)} (k_{p_s}^1, k_{p_s}^2) + O(|\rho|^4)$. For sufficiently small ρ this is incompatible with the hypothesis of noncollinearity of the lemma, and hence $\partial_c m_e(0, c_0)$ must be invertible.

Once a solution $c(v)$ of $m(v, c) = 0$ is established for $v \neq 0$, another argument using the statement of the implicit function theorem implies that $c(v)$ is locally real analytic away from the origin. Indeed, for $\rho \neq 0$ and for $e \in S_1 = \{v \in X_i : |v| = 1\}$ such that $(I_1(e), I_2(e)) \not\sim k_p$, the implicit function theorem implies that the solution $c(v) = C(\rho e)$ is real analytic in the parameter e. This concludes the proof of the lemma.

5. Analyticity of the Dirichlet–Neumann operator. The one remaining detail in our proof is the analyticity of the Dirichlet–Neumann operator. This was first established in the two dimensional setting by R. Coifman and Y. Meyer [4], and in the three dimensional setting by W. Craig, U. Schanz, and C. Sulem [7]. The result in general n dimensions was proven by D. Nicholls [19], as a generalization of the method of W. Craig, U. Schanz, and C. Sulem. All of these authors proved the following result in differing numbers of dimensions.

THEOREM 5.1. *Consider functions η such that $|\eta|_{L^\infty} < hR_0$, $|\eta|_{C^1} < R_0$, and $|\eta|_{C^{s+1}} < \infty$. There exists $R_0 > 0$ such that the Dirichlet–Neumann operator $G(\eta)$ is analytic in η in the neighborhood,*

$$\{\eta \mid \ |\eta|_{C^1} < R_0, |\eta|_{C^{s+1}} < \infty\}, \tag{5.1}$$

as a linear map in ξ from $W^{s+1,q} \to W^{s,q}$.

In this theorem C^s is the space of s times continuously differentiable functions, and $W^{s,q}$ is the L^q based Sobolev space of s times differentiable functions. In his thesis D. Nicholls [19] modified this theorem in two dimensions by only requiring $\eta \in W^{s+1,q}$, which is used to prove the result of this paper in two dimensions with

$\sigma = 0$. The approach of W. Craig, U. Schanz, and C. Sulem [7] was to perform estimates on the Dirichlet–Neumann operator by always placing C^s norms on η terms and $W^{s,q}$ norms on ξ terms. This produces a very clean result, which is not, however, useful for the applications in the present paper. In this section we replace the C^{s+1} derivative with a $W^{s+1,q}$ at the cost of a C^2 derivative on ξ. In this section we generalize this procedure to higher dimensions for use in the three dimensional result.

5.1. An exact implicit formula. The method of W. Craig, U. Schanz, and C. Sulem [7] begins with an exact implicit formula for the Dirichlet–Neumann operator. The formula is in terms of smoothing and singular integral operators which one can analyze using the theorem of M. Christ and J. Journé [3]. In this section we consider n dimensional real space $\mathbb{R}^n$ and will use the notation $x = (x', x_n) \in \mathbb{R}^{n-1} \times \mathbb{R}$ for a point in $\mathbb{R}^n$ rather than the (x, y) notation of previous sections.

THEOREM 5.2. *An exact implicit formula for the Dirichlet–Neumann operator in $n \geq 2$ dimensions is*

$$(I - B(\eta))\ G(\eta)\ \xi = |D_{x'}| \tanh(h\,|D_{x'}|)\ \xi + A(\eta)\ \xi, \tag{5.2}$$

where

$$A(\eta)\ \zeta = -\left(1 + e^{-2h|D_{x'}|}\right)^{-1} |D_{x'}|\, M(\eta)\ \zeta\ ,$$
$$B(\eta)\ \zeta = -\left(1 + e^{-2h|D_{x'}|}\right)^{-1} |D_{x'}|\, L(\eta)\ \zeta\ ,$$

and

$$M(\eta)\ \zeta = \int m_n(x', y')\zeta(y')\ dy'\ ,$$
$$L(\eta)\ \zeta = \int l_n(x', y')\zeta(y')\ dy'\ .$$

In two dimensions

$$\begin{aligned} m_2(x', y') = \frac{1}{\pi} \Bigg[& \frac{1}{(x' - y')^2} \left\{ \frac{(x' - y')\partial_{y'}\eta(y') - (\eta(x') - \eta(y'))}{1 + q_1^2} \right\} \\ & + \frac{1}{(x' - y')^2 + 4h^2} \left\{ \frac{(x' - y')\partial_{y'}\eta(y') + (\eta(x') + \eta(y'))}{1 + \frac{4h}{\sqrt{(x'-y')^2+4h^2}} q_2 + q_2^2} \right\} \\ & + \frac{2h}{(x' - y')^2 + 4h^2} \left\{ \frac{1}{1 + \frac{4h}{\sqrt{(x'-y')^2+4h^2}} q_2 + q_2^2} - 1 \right\} \Bigg] \end{aligned}$$

and

$$l_2(x', y') = -\frac{1}{2\pi} \left[\log(1 + q_1^2) + \log\left(1 + \frac{4h}{\sqrt{(x' - y')^2 + 4h^2}} q_2 + q_2^2 \right) \right].$$

In these formulae we set

$$q_1(x', y') = \frac{\eta(x') - \eta(y')}{|x' - y'|},$$
$$q_2(x', y') = \frac{\eta(x') + \eta(y')}{\sqrt{|x' - y'|^2 + 4h^2}}\ .$$

In $n \geq 3$ dimensions,

$$m_n(x',y') = \frac{2}{\omega_n}\left[\frac{1}{|x'-y'|^n}\left\{\frac{(x'-y')\cdot\nabla_{y'}\eta(y') - (\eta(x')-\eta(y'))}{(1+q_1^2)^{n/2}}\right\}\right.$$
$$+\frac{1}{\left(|x'-y'|^2+4h^2\right)^{\frac{n}{2}}}\left\{\frac{(x'-y')\cdot\nabla_{y'}\eta(y') + (\eta(x')+\eta(y'))}{\left(1+\frac{4h}{\sqrt{|x'-y'|^2+4h^2}}q_2+q_2^2\right)^{n/2}}\right\}$$
$$\left.+\frac{2h}{\left(|x'-y'|^2+4h^2\right)^{\frac{n}{2}}}\left\{\frac{1}{\left(1+\frac{4h}{\sqrt{|x'-y'|^2+4h^2}}q_2+q_2^2\right)^{n/2}}-1\right\}\right],$$

and

$$l_n(x',y') = \frac{2}{(n-2)\omega_n}\left[\frac{1}{|x'-y'|^{n-2}}\left\{\frac{1}{(1+q_1^2)^{(n-2)/2}}-1\right\}\right.$$
$$\left.+\frac{1}{\left(|x'-y'|^2+4h^2\right)^{(n-2)/2}}\left\{\frac{1}{\left(1+\frac{4h}{\sqrt{|x'-y'|^2+4h^2}}q_2+q_2^2\right)^{(n-2)/2}}-1\right\}\right],$$

where $\omega_n = 2\sqrt{\pi}^n/\Gamma(\frac{n}{2})$.

Proof. The proof of this result can be found in the paper of W. Craig, U. Schanz, and C. Sulem [7] for the case of three dimensions, or in the thesis of D. Nicholls [19] for the general case. The essence of the proof is that one can express solutions of Laplace's equation at a point in terms of the function itself and the fundamental solution. From here one takes appropriate derivatives, evaluates at the free surface, identifies the Dirichlet–Neumann operator, and then recognizes the linear terms as convolutions and writes them as $|D_{x'}|\tanh(h\,|D_{x'}|)$. In case that the depth h is infinite, the above Fourier multiplier is replaced by $|D_{x'}|$, the expression q_2 vanishes, and the subsequent analysis is similar but simpler. We will carry out the analysis for the case h finite in the remainder of this section. ∎

5.2. Proof of analyticity. The principal ingredients are the formulae of section 5.1 for the Dirichlet–Neumann operator and the theorems on singular and smoothing integral operators of section 5.3. The method we use is similar to the one employed by W. Craig, U. Schanz, and C. Sulem [7] and D. Nicholls [19]. We modify the approach by considering function spaces which are relevant to our current purpose. In particular, we will require only that $\eta \in W^{s+1,q}$ rather than C^{s+1} in the two dimensional setting, and $\eta \in W^{s+2,q}$ rather than C^{s+1} in the three dimensional setting, representing an improvement in the smoothness required of η—however, at the cost of the information and elegance of the proof of W. Craig, U. Schanz, and C. Sulem [7], in particular information involving $|\eta|_{C^1}$.

By investigating the analyticity properties of the operators $A(\eta)$ and $B(\eta)$, along with the analyticity of the operator $(I - B(\eta))$, we will show that the Dirichlet–Neumann operator itself is analytic. We begin by considering the operator $B(\eta)$.

THEOREM 5.3. *In the setting of the two dimensional water wave problem ($n = 2$), if $1 < q < \infty$ and $s > \max(\frac{1}{q}, 2)$, then $B(\eta)$ satisfies the estimate*

$$\|B(\eta)\ \xi\|_{W^{s,q}} < C\,\|\eta\|^2_{W^{s+1,q}}\,\|\xi\|_{W^{s,q}}\,. \tag{5.3}$$

Furthermore, the operator $B(\eta)$ is analytic as a function of η on the space $W^{s,q}$. In the setting of the three dimensional water wave problem ($n = 3$), if $1 < q < \infty$ and $s > \max(\frac{2}{q}, 3)$, then $B(\eta)$ satisfies the estimate

$$\|B(\eta)\ \xi\|_{W^{s,q}} < C\,\|\eta\|^2_{W^{s+2,q}}\,\|\xi\|_{W^{s,q}}\,. \tag{5.4}$$

Furthermore, the operator $B(\eta)$ is analytic as a function of η on the space $W^{s,q}$.

The proof of this theorem depends upon the following two lemmas which are proven in W. Craig, U. Schanz, and C. Sulem [7] and D. Nicholls [19].

LEMMA 5.4. *The operator $\left(1 + e^{-2h|D_{x'}|}\right)^{-1}$ and the Riesz potential $R_j(D_{x'}) = i\frac{D_{x'_j}}{|D_{x'}|}$ are bounded on $W^{s,q}$ for $1 < q < \infty$ and $s \geq 0$.*

LEMMA 5.5. *Given functions $f \in C^s$ and $g \in W^{s,q}$ the following interpolation identity holds for some constant $K(s)$:*

$$\|fg\|_{W^{s,q}} \leq K(s)\left[|f|_{L^\infty}\,\|g\|_{W^{s,q}} + |f|_{C^s}\,\|g\|_{L^q}\right]. \tag{5.5}$$

We may now proceed with the proof of Theorem 5.3. Since

$$|D_{x'}| = -\sum_{j=1}^{n-1} R_j(D_{x'})\partial_{x'_j},$$

and we have Lemma 5.4, we need only consider two types of integral operators, $\partial_{x'_j}$ applied to

$$P_2 = \int -\frac{1}{2\pi}\log(1 + q_1^2)\ \xi(y')\ dy',$$
$$Q_2 = \int -\frac{1}{2\pi}\log(1 + \kappa_h q_2 + q_2^2)\ \xi(y')\ dy'$$

in two dimensions and

$$P_n = \int \frac{2}{(n-2)\omega_n}\frac{1}{|x' - y'|^{n-2}}\left\{\frac{1}{(1+q_1^2)^{(n-2)/2}} - 1\right\}\xi(y')\ dy'$$
$$Q_n = \int \frac{2}{(n-2)\omega_n}\frac{1}{\left(|x' - y'|^2 + 4h^2\right)^{(n-2)/2}}$$
$$\times\left\{\frac{1}{(1+\kappa_h q_2 + q_2^2)^{(n-2)/2}} - 1\right\}\xi(y')\ dy'$$

in $n \geq 3$ dimensions, where $\kappa_h = 4h/\sqrt{|x' - y'|^2 + 4h^2}$. Applying the derivative operator to the above quantities results in the following formulae:

$$\partial_{x'}P_2 = -\frac{1}{\pi}\int \frac{q_1}{1+q_1^2}\left[\frac{\partial_{x'}\eta(x')}{|x' - y'|} - \frac{(\eta(x') - \eta(y'))(x' - y')}{|x' - y'|^3}\right]\xi(y')\ dy',$$

$$\partial_{x'}Q_2 = -\frac{1}{2\pi}\int \frac{1}{1+\kappa_h q_2+q_2^2}\left[-q_2\frac{4h(x'-y')}{\left(|x'-y'|^2+4h^2\right)^{3/2}}\right.$$
$$\left.+(\kappa_h+2q_2)\left\{\frac{\partial_{x'}\eta(x')}{\left(|x'-y'|^2+4h^2\right)^{1/2}} -\frac{(\eta(x')+\eta(y'))(x'-y')}{\left(|x'-y'|^2+4h^2\right)^{3/2}}\right\}\right]\xi(y')\,dy'$$

in two dimensions and

$$\partial_{x'_j}P_n = -\frac{2}{\omega_n}\int\frac{x'_j-y'_j}{|x'-y'|^n}\left\{\frac{1}{(1+q_1^2)^{(n-2)/2}}-1\right\}\xi(y')\,dy'$$
$$-\frac{2}{\omega_n}\int\frac{1}{|x'-y'|^{n-2}}\frac{q_1}{(1+q_1^2)^{n/2}}\left[\frac{\partial_{x'_j}\eta(x')}{|x'-y'|}\right.$$
$$\left.-\frac{(\eta(x')-\eta(y'))(x'_j-y'_j)}{|x'-y'|^3}\right]\xi(y')\,dy'\ ,$$
$$\partial_{x'_j}Q_n = \frac{2}{\omega_n}\int\frac{x'_j-y'_j}{\left(|x'-y'|^2+4h^2\right)^{n/2}}\left\{\frac{1}{(1+\kappa_h q_2+q_2^2)^{(n-2)/2}}-1\right\}\xi(y')\,dy'$$
$$+\frac{2}{(n-2)\omega_n}\int\frac{1}{\left(|x'-y'|^2+4h^2\right)^{(n-2)/2}}\frac{1}{(1+\kappa_h q_2+q_2^2)^{n/2}}$$
$$\times\left\{-q_2\frac{4h(x'_j-y'_j)}{\left(|x'-y'|^2+4h^2\right)^{3/2}}\right.$$
$$\left.+(\kappa_h+2q_2)\cdot\left[\frac{\partial_{x'_j}\eta(x')}{\left(|x'-y'|^2+4h^2\right)^{1/2}}-\frac{(\eta(x')+\eta(y'))(x'_k-y'_k)}{\left(|x'-y'|^2+4h^2\right)^{3/2}}\right]\right\}\xi(y')\,dy'$$

in $n\geq 3$ dimensions. It is not difficult to see that the singular and smoothing integral operator theorems of section 5.3 apply to the above operators if we keep in mind that $d=n-1$. Indeed, in the seven principal terms appearing in the expression for the operator $\partial_{x'_j}Q_n$ and in the notation of Theorem 5.21, we have $(p,\rho,\lambda)=(1,n-2,1)$, $(2,n-2,0)$, $(1,n,0)$, $(0,n-1,1)$, $(1,n-1,0)$, $(1,n,0)$, and $(0,n,1)$. It is also useful to use Lemma 5.5 to pull out functions which depend on x' from the y' integrals, and interpolate.

The goal is not to show that $B(\eta)$ is analytic but rather to show that $(I-B(\eta))^{-1}$ is analytic. With this in mind we make the following estimate concerning powers of the operator $B(\eta)$ and then conclude analyticity of $(I-B(\eta))^{-1}$.

COROLLARY 5.6. *In the setting of the two dimensional water wave problem ($n=2$), if $1<q<\infty$ and $s>\max(\frac{1}{q},2)$, then the powers of $B(\eta)$ satisfy the following estimate:*

$$\left\|B(\eta)^j\ \xi\right\|_{W^{s,q}} < C^j\left\|\eta\right\|^{2j}_{W^{s+1,q}}\left\|\xi\right\|_{W^{s,q}}. \tag{5.6}$$

Thus, for $\|\eta\|_{W^{s+1,q}}$ small enough, the operator $(I-B(\eta))^{-1}$ exists, satisfies the

estimate

$$\left\|(I - B(\eta))^{-1}\ \xi\right\|_{W^{s,q}} < 1 + \tilde{C}\,\|\xi\|_{W^{s,q}}\,, \tag{5.7}$$

and is analytic as a function of η *on* $W^{s,q}$*. In the setting of the three dimensional water wave problem* ($n = 3$)*, if* $1 < q < \infty$ *and* $s > \max(\frac{2}{q}, 3)$*, then the powers of* $B(\eta)$ *satisfy the following estimate:*

$$\left\|B(\eta)^j\ \xi\right\|_{W^{s,q}} < C^j\,\|\eta\|^{2j}_{W^{s+2,q}}\,\|\xi\|_{W^{s,q}}\,. \tag{5.8}$$

Thus, for $\|\eta\|_{W^{s+2,q}}$ *small enough, the operator* $(I - B(\eta))^{-1}$ *exists, satisfies the estimate*

$$\left\|(I - B(\eta))^{-1}\ \xi\right\|_{W^{s,q}} < 1 + \tilde{C}\,\|\xi\|_{W^{s,q}}\,, \tag{5.9}$$

and is analytic as a function of η *on* $W^{s,q}$*.*

Proof. We restrict to the case of $n = 2$ as the $n = 3$ case is virtually identical. We already have the estimate

$$\|B(\eta)\ \xi\|_{W^{s,q}} < C\,\|\eta\|^2_{W^{s+1,q}}\,\|\xi\|_{W^{s,q}}\,,$$

and we now proceed via induction. The case $j = 1$ is clearly true, so we assume that the corollary is true for j and analyze $(j + 1)$,

$$\begin{aligned}\left\|B(\eta)^{j+1}\ \xi\right\|_{W^{s,q}} &= \left\|B(\eta)^j B(\eta)\ \xi\right\|_{W^{s,q}} \\ &< C^j\,\|\eta\|^{2j}_{W^{s+1,q}}\,\|B(\eta)\ \xi\|_{W^{s,q}} \\ &< C^{j+1}\,\|\eta\|^{2(j+1)}_{W^{s+1,q}}\,\|\xi\|_{W^{s,q}}\,.\end{aligned}$$

Therefore, the estimate on the jth power holds true. To compute $(I - B(\eta))^{-1}$ we use the Neumann series

$$(I - B(\eta))^{-1} = I + \sum_{j=1}^{\infty} B^j(\eta)$$

and note that it converges in the radius of convergence of the series

$$\sum_{j=1}^{\infty} C^j\,\|\eta\|^{2j}_{W^{s+1,q}}\,.$$

In other words, the Neumann series converges when $\|\eta\|_{W^{s+1,q}} < 1/\sqrt{C}$, and we are done. ∎

We now turn our attention to the operators $A(\eta)$. While the operators $B(\eta)$ map $W^{s,q}$ into $W^{s,q}$ and thereby preserve derivatives, we should expect the operators $A(\eta)$ to map $W^{s+1,q}$ into $W^{s,q}$. This is incorporated into our proof by using a variation of integration by parts, given below, to remove all $\partial_{y'}$ derivatives from $\eta(y')$ terms and place them elsewhere. We give the version of integration by parts in the following lemma proven in W. Craig, U. Schanz, and C. Sulem [7] and D. Nicholls [19].

LEMMA 5.7. *Consider* $x, y \in \mathbb{R}^n$*,* $R(q_1)$ *an odd continuous function of* q_1*, and* $\eta \in C^1(\mathbb{R}^n)$*; then we have that*

$$\int \frac{x - y}{|x - y|^n} \cdot \nabla_y (R(q_1))\xi(y)\ dy = -\int R(q_1)\frac{x - y}{|x - y|^n} \cdot \nabla_y \xi(y)\ dy. \tag{5.10}$$

We are can now prove an estimate on the operator $A(\eta)$ which is given in the following theorem.

THEOREM 5.8. *In the setting of the two dimensional water wave problem ($n = 2$), if $1 < q < \infty$ and $s > \max(\frac{1}{q}, 2)$, then $A(\eta)$ satisfies the estimate*

$$\|A(\eta)\ \xi\|_{W^{s,q}} < C\,\|\eta\|^2_{W^{s+1,q}}\,\|\xi\|_{W^{s+1,q}}\,. \tag{5.11}$$

Furthermore, the operator $A(\eta)$ is analytic as a function of η from the space $W^{s+1,q}$ to the space $W^{s,q}$. In the setting of the three dimensional water wave problem ($n = 3$), if $1 < q < \infty$ and $s > \max(\frac{2}{q}, 3)$, then $A(\eta)$ satisfies the estimate

$$\|A(\eta)\ \xi\|_{W^{s,q}} < C\,\|\eta\|^2_{W^{s+2,q}}\,\|\xi\|_{W^{s+1,q}}\,. \tag{5.12}$$

Furthermore, the operator $A(\eta)$ is analytic as a function of η from the space $W^{s+1,q}$ to the space $W^{s,q}$.

Proof. As before, since

$$|D_{x'}| = -\sum_{j=1}^{n-1} R_j(D_{x'})\partial_{x'_j}$$

is the Riesz potential given in Lemma 5.4, we need consider only two types of integral operators, $\partial_{x'_j}$ applied to

$$\begin{aligned}
P_2 &= \int \frac{1}{\pi}\frac{1}{(x'-y')^2}\left\{\frac{(x'-y')\partial_{y'}\eta(y') - (\eta(x')-\eta(y'))}{1+q_1^2}\right\}\xi(y')\,dy',\\
Q_2 &= \int \frac{1}{\pi}\frac{1}{(x'-y')^2+4h^2}\left\{\frac{(x'-y')\partial_{y'}\eta(y')}{1+\kappa_h q_2+q_2^2}\right.\\
&\quad\left.+\frac{(\eta(x')+\eta(y'))}{1+\kappa_h q_2+q_2^2}\right\}\xi(y')\,dy',\\
R_2 &= \int \frac{1}{\pi}\frac{2h}{(x'-y')^2+4h^2}\left\{\frac{1}{1+\kappa_h q_2+q_2^2}-1\right\}\xi(y')\,dy'
\end{aligned}$$

in two dimensions and

$$\begin{aligned}
P_n &= \int \frac{2}{\omega_n}\frac{1}{|x'-y'|^n}\left\{\frac{(x'-y')\cdot\nabla_{y'}\eta(y') - (\eta(x')-\eta(y'))}{(1+q_1^2)^{n/2}}\right\}\xi(y')\,dy',\\
Q_n &= \int \frac{2}{\omega_n}\frac{1}{\left(|x'-y'|^2+4h^2\right)^{n/2}}\left\{\frac{(x'-y')\cdot\nabla_{y'}\eta(y')}{(1+\kappa_h q_2+q_2^2)^{n/2}}\right.\\
&\quad\left.+\frac{(\eta(x')+\eta(y'))}{(1+\kappa_h q_2+q_2^2)^{n/2}}\right\}\xi(y')\,dy',\\
R_n &= \int \frac{2}{\omega_n}\frac{2h}{\left(|x'-y'|^2+4h^2\right)^{n/2}}\left\{\frac{1}{(1+\kappa_h q_2+q_2^2)^{n/2}}-1\right\}\xi(y')\,dy'
\end{aligned}$$

in $n \geq 3$ dimensions. All we need to do is apply the differential operator $\partial_{x'_j}$ to each of these integrals, use Lemma 5.7 wherever appropriate, and then note that the singular

and smoothing integral operator theorems of section 5.3 can be used. Keeping in mind the facts that

$$\partial_{x_j'} q_1(x',y') = \frac{\partial_{x_j'}\eta(x')}{|x'-y'|} - \frac{(\eta(x')-\eta(y'))(x_j'-y_j')}{|x'-y'|^3},$$
$$\partial_{x_j'} q_2(x',y') = \frac{\partial_{x_j'}\eta(x')}{\left(|x'-y'|^2+4h^2\right)^{1/2}} - \frac{(\eta(x')+\eta(y'))(x_j'-y_j')}{\left(|x'-y'|^2+4h^2\right)^{3/2}},$$
$$\partial_{x_j'}\kappa_h(x',y') = -\frac{4h(x_j'-y_j')}{\left(|x'-y'|^2+4h^2\right)^{3/2}},$$

we compute the derivatives in two dimensions as

$$\begin{aligned}
\partial_{x'}P_2 = &-\frac{2}{\pi}\int \frac{1}{(x'-y')^3}\left\{\frac{(x'-y')\cdot\nabla_{y'}\eta(y')-(\eta(x')-\eta(y'))}{(1+q_1^2)^{n/2}}\right\}\xi(y')\,dy' \\
&+\frac{1}{\pi}\int \frac{1}{(x'-y')^2}\left\{\frac{\partial_{y'}\eta(y')-\partial_{x'}\eta(x')}{1+q_1^2}\right. \\
&\left.- \frac{\left[(x'-y')\partial_{y'}\eta(y')-(\eta(x')-\eta(y'))\right]2q_1\partial_{x'}q_1}{\left(1+q_1^2\right)^2}\right\}\xi(y')\,dy', \\
\partial_{x'}Q_2 = &-\frac{2}{\pi}\int \frac{x'-y'}{\left((x'-y')^2+4h^2\right)^2}\left\{\frac{(x'-y')\partial_{y'}\eta(y')+(\eta(x')+\eta(y'))}{1+\kappa_h q_2+q_2^2}\right\}\xi(y')\,dy' \\
&+\frac{1}{\pi}\int \frac{1}{(x'-y')^2+4h^2}\left\{\frac{\partial_{y'}\eta(y')+\partial_{x'}\eta(x')}{1+\kappa_h q_2+q_2^2}\right. \\
&\left.- \frac{\left[(x'-y')\partial_{y'}\eta(y')+(\eta(x')+\eta(y'))\right](q_2\partial_{x'}\kappa_h+(\kappa_h+2q_2)\partial_{x'}q_2)}{(1+\kappa_h q_2+q_2^2)^2}\right\}\xi(y')\,dy', \\
\partial_{x'}R_2 = &-\frac{2}{\pi}\int \frac{2h(x'-y')}{\left((x'-y')^2+4h^2\right)^2}\left\{\frac{1}{1+\kappa_h q_2+q_2^2}-1\right\}\xi(y')\,dy' \\
&-\frac{1}{\pi}\int \frac{2h}{(x'-y')^2+4h^2}\frac{(q_2\partial_{x'}\kappa_h+(\kappa_h+2q_2)\partial_{x'}q_2)}{\left(1+\kappa_h q_2+q_2^2\right)^2}
\end{aligned}$$

and in $n=3$ dimensions as

$$\begin{aligned}
\partial_{x_j'}P_n = &\frac{2}{\omega_n}\int \frac{x_j'-y_j'}{|x'-y'|^{n+2}}\left\{\frac{(x'-y')\cdot\nabla_{y'}\eta(y')-(\eta(x')-\eta(y'))}{(1+q_1^2)^{n/2}}\right\}\xi(y')\,dy' \\
&+\frac{2}{\omega_n}\int \frac{1}{|x'-y'|^n}\left\{\frac{\partial_{y_j'}\eta(y')-\partial_{x_j'}\eta(x')}{\left(1+q_1^2\right)^{n/2}}\right. \\
&\left.-\frac{\left[(x'-y')\cdot\nabla_{y'}\eta(y')-(\eta(x')-\eta(y'))\right](nq_1\partial_{x_j'}q_1)}{\left(1+q_1^2\right)^n}\right\}\xi(y')\,dy', \\
\partial_{x_j'}Q_n = &-\frac{2n}{\omega_n}\int \frac{x_j'-y_j'}{\left(|x'-y'|^2+4h^2\right)^{(n+2)/2}} \\
&\times\left\{\frac{(x'-y')\cdot\nabla_{y'}\eta(y')+(\eta(x')+\eta(y'))}{\left(1+\kappa_h q_2+q_2^2\right)^{n/2}}\right\}\xi(y')\,dy'
\end{aligned}$$

$$+\frac{2}{\omega_n}\int \frac{1}{\left(|x'-y'|^2+4h^2\right)^{n/2}}\left\{\frac{\partial_{y_j'}\eta(y')+\partial_{x_j'}\eta(x')}{(1+\kappa_h q_2+q_2^2)^{n/2}}\right.$$
$$\left.-\frac{[(x'-y')\cdot\nabla_{y'}\eta(y')+(\eta(x')+\eta(y'))]\,(q_2\partial_{x_j'}\kappa_h+(\kappa_h+2q_2)\partial_{x_j'}q_2)}{(1+\kappa_h q_2+q_2^2)^{n}}\right\}\xi(y')\,dy',$$

$$\partial_{x_j'}R_n=-\frac{2n}{\omega_n}\int\frac{2h(x_j'-y_j')}{\left(|x'-y'|^2+4h^2\right)^{(n+2)/2}}\left\{\frac{1}{(1+\kappa_h q_2+q_2^2)^{n/2}}-1\right\}\xi(y')\,dy'$$
$$-\frac{n}{\omega_n}\int\frac{2h}{\left(|x'-y'|^2+4h^2\right)^{n/2}}\frac{q_2\partial_{x_j'}\kappa_h+(\kappa_h+2q_2)\partial_{x_j'}q_2}{(1+\kappa_h q_2+q_2^2)^{(n+2)/2}}\xi(y')\,dy'.$$

As noted before, we can now use the theorems of section 5.3 in combination with Lemma 5.7 to arrive at the conclusion of the theorem. □

Now that we have Corollary 5.6 and Theorem 5.8, we can finally state and prove the theorem regarding analyticity of the Dirichlet–Neumann operator.

THEOREM 5.9. *In the setting of the two dimensional water wave problem ($n=2$), if $1<q<\infty$ and $s>\max(\frac{1}{q},2)$, then the Dirichlet–Neumann operator $G(\eta)\,\xi$ is analytic as a function of $\eta\in W^{s+1,q}$, as a bounded linear operator from $W^{s+1,q}$ to $W^{s,q}$. In the setting of the three dimensional water wave problem ($n=3$), if $1<q<\infty$ and $s>\max(\frac{2}{q},3)$, then the Dirichlet–Neumann operator $G(\eta)\,\xi$ is analytic as a function of $\eta\in W^{s+2,q}$, as a bounded linear operator from $W^{s+1,q}$ to $W^{s,q}$.*

Proof. Theorem 5.2 gives the exact implicit formula

$$(I-B(\eta))G(\eta)\xi=|D_{x'}|\tanh(h\,|D_{x'}|)\xi+A(\eta)\xi,$$

and thus, since $(I-B(\eta))$ is boundedly invertible, we can write

$$G(\eta)\xi=(I-B(\eta))^{-1}\,|D_{x'}|\tanh(h\,|D_{x'}|)\xi+(I-B(\eta))^{-1}A(\eta)\xi.$$

Since Corollary 5.6 and Theorem 5.8 give us analyticity and appropriate boundedness of relevant operators we have the statement of the theorem. □

5.3. Singular and smoothing integral operators. In this section we briefly outline the statement and proofs of theorems concerning the boundedness properties of certain singular and smoothing integral operators relevant to the proof the the analyticity of the Dirichlet–Neumann operator outlined above. The bulk of these theorems were first presented in W. Craig, U. Schanz, and C. Sulem [7] and U. Schanz [25] in general n dimensions for use in the analyticity proof of the Dirichlet–Neumann operator in three dimensions where $\eta\in C^{s+1}$ and $\xi\in W^{s,q}$. The others were first presented by D. Nicholls [19] in one dimension for use in the proof of the analyticity of the Dirichlet–Neumann operator in two dimensions which only required that $\eta,\xi\in W^{s,q}$ for $s>\frac{d}{q}$. Here we extend the program of D. Nicholls [19] by proving his result in d dimensions, and further reducing the smoothness required of η and ξ.

Let $x,y\in\mathbb{R}^d$, and consider functions $\eta:\mathbb{R}^d\to\mathbb{R}$ and $\xi:\mathbb{R}^d\to\mathbb{R}$. Our goal is to study two classes of integral operators. The first is singular and has the general form

$$C_p(\eta)\;\xi(x)=\int k(x-y)c_p(q_1)\xi(y)\,dy,$$

where $k(x-y)$ is a convolution kernel of Calderón–Zygmund type satisfying standard estimates outlined below. Recall that $q_1(\eta; x, y) = (\eta(x) - \eta(y))/|x-y|$, and consider $c_p : \mathbb{R} \to \mathbb{R}$ which is analytic in the interval $|z| < R_0$ such that $c_p(z) = \mathcal{O}(|z|^p)$ for $|z|$ small. The second class of operators is smoothing and has the general form

$$C_{p,h}(\eta)\ \xi(x) = \int K_h(x-y) c_{p,h}(q_2, \kappa_h) \xi(y)\ dy,$$

where

$$K_h(x) = \frac{1}{\left(|x|^2 + 4h^2\right)^{\rho/2}} \prod_{l=1}^{d} \left(\frac{x_l}{\left(|x|^2 + 4h^2\right)^{1/2}} \right)^{\beta_l},$$

with $\beta = \sum_{l=1}^{d} \beta_l$,

$$q_2(\eta; x, y) = \frac{\eta(x) + \eta(y)}{\left(|x-y|^2 + 4h^2\right)^{1/2}}, \quad \kappa_h(x,y) = \frac{4h}{\left(|x-y|^2 + 4h^2\right)^{1/2}}.$$

We consider $c_{p,h}(z, w) : \mathbb{R}^2 \to \mathbb{R}$ which is analytic for $\{|z| < R_0, |w| < 2\}$ such that $c_{p,h}(z,w) = \mathcal{O}(|z|^p |w|^\lambda)$ for $|z|$ and $|w|$ small. We will require that $p + \rho + \lambda > d$. We will establish Sobolev estimates for the operators C_p and $C_{p,h}$ in the following theorems. The first, concerning bounds on C_p, requires the use of a deep theorem of M. Christ and J. Journé [3] on L^q bounds for Calderón–Zygmund commutators. The theorem concerning bounds on the operators $C_{p,h}$ requires nothing more than careful estimates as the operator is smoothing rather than singular.

To place ourselves in the setting for the theorem of M. Christ and J. Journé [3], we define the standard estimates.

DEFINITION 5.10. *A kernel K on $\mathbb{R}^d$ is said to satisfy standard estimates if there exist $\delta > 0$ and $c_K < \infty$ such that for all distinct $x, y \in \mathbb{R}^d$ and all $z \in \mathbb{R}^d$ such that $|x - z| \le \frac{|x-y|}{2}$,*

(i) $|K(x,y)| \le c_K |x-y|^{-d}$;
(ii) $|K(x,y) - K(z,y)| \le c_K (\frac{|x-z|}{|x-y|})^\delta |x-y|^{-d}$;
(iii) $|K(y,x) - K(y,z)| \le c_K (\frac{|x-z|}{|x-y|})^\delta |x-y|^{-d}$.

With this definition in hand we state the following theorem.

THEOREM 5.11 (M. Christ and J. Journé [3, Theorem 4]). *Consider the singular integral operator with kernel*

$$L(x-y) \prod_{j=1}^{m} \left(\int_0^1 b_j(tx + (1-t)y)\ dt \right), \tag{5.13}$$

where each $b_j \in L^\infty$ and L satisfies the standard estimates. Then for $0 < q < \infty$,

$$\left\| \int L(x-y) \prod_{j=1}^{m} \left(\int_0^1 b_j(tx + (1-t)y)\ dt \right) \xi(y)\ dy \right\|_{L^q} \le c_q c_L m^N \left(\prod_{j=1}^{m} |b_j|_{L^\infty} \right) \|\xi\|_{L^q}. \tag{5.14}$$

One may take $N = 2 + \delta_L$, for δ_L and c_L which appear in the standard estimates.

The first step in analyzing the boundedness of the operators $C_p(\eta)$ is to first consider the simplified operator

$$S_p(\eta_1, \ldots, \eta_p)\ \xi(x) = \int K(x-y) \left(\prod_{j=1}^{p} q_1(\eta_j) \right) \xi(y)\ dy,$$

where K is a Calderón–Zygmund kernel which satisfies standard estimates. There are two related theorems which one can prove concerning this operator. The first is due to W. Craig, U. Schanz, and C. Sulem [7].

THEOREM 5.12 (W. Craig, U. Schanz, and C. Sulem [7]). *Let $\eta_1, \ldots, \eta_p \in C^1$, then the singular integral operator $S_p(\eta_1, \ldots, \eta_p)$ is bounded on L^q and satisfies*

$$\|S_p(\eta_1, \ldots, \eta_p)\ \xi\|_{L^q} \leq C_0 p^M \left(\prod_{j=1}^{p} |\eta_j|_{C^1} \right) \|\xi\|_{L^q}\,, \tag{5.15}$$

with exponent $M = 3 + \min(\delta_K, 1)$.

The second is an extension to general d dimensions of a result proven by D. Nicholls [19].

THEOREM 5.13. *Let $1 \leq r \leq p$, $\eta_j \in C^2$ for $j \neq r$, $\xi \in C^1 \bigcap W^{1,q}$, and $\eta_r \in L^\infty \bigcap L^q$. Consider K of the form*

$$K(x,y) = \begin{cases} \frac{1}{|x-y|^d}, \\ \frac{x-y}{|x-y|^{d+1}}. \end{cases} \tag{5.16}$$

Then the singular integral operator $S_p(\eta_1, \ldots, \eta_p)$ is bounded on L^q and satisfies

$$\begin{aligned} \|S_p(\eta_1, \ldots, \eta_p)\ \xi\|_{L^q} \leq C \Bigg\{ & |\eta_r|_{L^\infty} \left(\prod_{j=1, j\neq r}^{p} |\eta_j|_{C^1} \right) \|\partial_y \xi\|_{L^q} \\ & + |\eta_r|_{L^\infty} \sum_{s=1, s\neq r}^{p} \left(\prod_{j=1, j\neq r, s}^{p} |\eta_j|_{C^1} \right) |\eta_s|_{C^2} \|\xi\|_{L^q} \\ & + \left(\prod_{j=1, j\neq r}^{p} |\eta_j|_{C^1} \right) |\xi|_{C^1} \|\eta_r\|_{L^q} \\ & + |\xi|_{L^\infty} \left(\prod_{j=1, j\neq r}^{p} |\eta_j|_{C^1} \right) \|\partial_y \eta_r\|_{L^q} \\ & + |\xi|_{L^\infty} \sum_{s=1, s\neq r}^{p} \left(\prod_{j=1, j\neq r, s}^{p} |\eta_j|_{C^1} \right) |\eta_s|_{C^2} \|\eta_r\|_{L^q} \Bigg\}. \end{aligned} \tag{5.17}$$

The proofs of both of these results rely on the following lemmas. We state without proof the first two (see [19]) which are used in the proof of Theorem 5.12.

LEMMA 5.14. *Suppose that $x \neq y$. If $f_1, \ldots, f_p \in C^1$, then*

$$\prod_{j=1}^{p} q_1(f_j) = \sum_{l \in \mathcal{L}} \left(\prod_{k=1}^{p} \frac{(x-y)_{l_k}}{|x-y|} \right) \left(\int_0^1 \partial_{x_{l_k}} f_j(tx + (1-t)y)\ dt \right), \tag{5.18}$$

where $\mathcal{L}$ is the set of all integer p-tuples $(l_1, \ldots, l_p)$ such that $1 \le l_1, \ldots, l_p \le d$.

LEMMA 5.15. *If $k(x,y)$ is a Calderón–Zygmund kernel satisfying standard estimates, then the kernel of the form*

$$\zeta(x,y) = k(x,y) \prod_{k=1}^{p} \frac{(x-y)_{l_k}}{|x-y|}, \tag{5.19}$$

where $1 \le l_k \le d$, also satisfies standard estimates.

The following two lemmas are the analogues of the above two in the case that a y derivative is applied to $q_1(f)$. The proof of Lemma 5.16 is significantly different from the proof of Lemma 5.14 so we present it here. However, the proof of Lemma 5.17 is sufficiently close to that of Lemma 5.15 that we omit it.

LEMMA 5.16. *Suppose that $x \neq y$. If $f \in C^2$, then*

$$\begin{aligned} \frac{x-y}{|x-y|} \cdot \nabla_y q_1(f) &= -\frac{x-y}{|x-y|^2} \cdot \nabla_y f(y) + \frac{f(x)-f(y)}{|x-y|^2} \\ &= \frac{1}{|x-y|^2} \sum_{k=1}^{d} (x-y)_k \sum_{j=1}^{d} (x-y)_j \int_0^1 b_{j,k}(x,y,t)\, dt, \end{aligned} \tag{5.20}$$

where

$$b_{j,k} = \int_0^1 \tau \partial_{y_j} \partial_{y_k} f(t[\tau x + (1-\tau)y] + (1-t)y)\, d\tau \tag{5.21}$$

and $b_{j,k} \in L^\infty$.

Proof. The first line of (5.20) is realized by a simple calculation of the derivative of $q_1(f)$ with respect to y. For the second line we begin with the fundamental theorem of calculus which states that for $f \in C^1$,

$$f(b) - f(a) = \sum_{j=1}^{d} \int_0^1 (b-a)_j \partial_{u_j} f(tb + (1-t)a)\, dt.$$

Provided that $f \in C^2$ this can also be done for ∇f resulting in the following:

$$\nabla_u f(b) - \nabla_u f(a) = \sum_{k=1}^{d} \hat{e}_k \sum_{j=1}^{d} \int_0^1 (b-a)_j \partial_{u_j} \partial_{u_k} f(tb + (1-t)a)\, dt,$$

where $\hat{e}_k$ is the kth unit vector. We now set $b = \tau x + (1-\tau)y$ and $a = y$, dot with $(x-y)$, and integrate in τ from 0 to 1. Using the fundamental theorem of calculus on the $\nabla_u f(b)$ term we arrive at

$$\begin{aligned} &(f(x) - f(y)) - (x-y) \cdot \nabla_u f(y) \\ &\qquad = \sum_{k=1}^{d} (x-y)_k \sum_{j=1}^{d} (x-y)_j \int_0^1 \tau \int_0^1 \partial_{u_j} \partial_{u_k} f(t[\tau x + (1-\tau)y] + (1-t)y)\, dt\, d\tau. \end{aligned}$$

We now multiply both sides by $\frac{1}{|x-y|^2}$ and the theorem is proven. That $b_{j,k}(x,y,t) \in L^\infty$ is due to the facts that $f \in C^2$ and $\tau \in [0,1]$. ∎

LEMMA 5.17. *A Calderón–Zygmund kernel of the type*

$$\zeta(x,y) = k(x,y)\frac{1}{|x-y|^2}\sum_{r,s}(x-y)_r(x-y)_s\prod_{k=1}^{p}\frac{(x-y)_{l_k}}{|x-y|}, \tag{5.22}$$

where $1 \le r,s,l_k \le d$ *and* $k(x,y)$ *is a Calderón–Zygmund kernel satisfying standard estimates, also satisfies standard estimates.*

At this point we can prove both Theorem 5.12 and Theorem 5.13. We begin with Theorem 5.12.

Proof of Theorem 5.12. The main idea is to use the theorem of M. Christ and J. Journé [3] on the operator $S_p(\eta_1,\ldots,\eta_p)$. Define

$$L(x,y) = K(x,y)\prod_{j=1}^{p}\frac{(x-y)_{l_k}}{|x-y|},$$

with K as defined in the statement of the theorem and $1 \le l_k \le d$ as in Lemma 5.14. By Lemma 5.14 we can write the kernel of S_p

$$K(x,y)\prod_{j=1}^{p} q_1(\eta_j)$$

as a finite sum of terms of the form

$$L(x,y)\prod_{j=1}^{p}\int_0^1 \partial_{x_{l_k}}\eta_j(tx+(1-t)y)\,dt.$$

By Lemma 5.15 we know that $L(x,y)$ satisfies standard estimates with $c_L = 3pc_K$ and $\delta_L = \min(\delta_K,1)$. Therefore we can apply the theorem of M. Christ and J. Journé above to each term

$$\left\|\int L(x,y)\prod_{j=1}^{p}\left(\int_0^1 \partial_{x_{l_k}}\eta_j(tx+(1-t)y)\,dt\right)\xi(y)\,dy\right\|_{L^q} \le c_L p^N\left(\prod_{j=1}^{p}\left|\partial_{x_{l_k}}\eta_j\right|_{L^\infty}\right)\|\xi\|_{L^q}.$$

By the definition of the C^1 norm the theorem is proven. □

The proof of Theorem 5.13 is in many ways a corollary of Theorem 5.12 and is presented below.

Proof of Theorem 5.13. The main idea is to use the theorem of M. Christ and J. Journé [3] on the operator S_p. We begin with the calculation

$$\begin{aligned} S_p\xi(x) &= \int K(x,y)\left(\prod_{j=1}^{p} q_1(\eta_j)\right)\xi(y)\,dy \\ &= \int K(x,y)\left(\prod_{j=1,j\neq r}^{p} q_1(\eta_j)\right)\frac{\eta_r(x)\xi(y)-\eta_r(y)\xi(y)}{|x-y|}\,dy \end{aligned}$$

$$
\begin{aligned}
&= \int K(x,y) \left(\prod_{j=1, j\neq r}^{p} q_1(\eta_j) \right) q_1(\xi)\eta_r(y)\, dy \\
&\quad + \int K(x,y) \left(\prod_{j=1, j\neq r}^{p} q_1(\eta_j) \right) \frac{\eta_r(x)\xi(y)}{|x-y|}\, dy \\
&\quad - \int K(x,y) \left(\prod_{j=1, j\neq r}^{p} q_1(\eta_j) \right) \frac{\eta_r(y)\xi(x)}{|x-y|}\, dy \\
&= I_1 + I_2 + I_3.
\end{aligned}
$$

Using Theorem 5.12 with the roles of η_r and ξ interchanged we can estimate the integral I_1 in appropriate fashion. We note that this estimate requires that $\eta_j, \xi \in C^1$ for $j \neq r$, and $\eta_r \in L^q$. The two integrals I_2 and I_3 can be handled in the same way as one another, again with the roles of η_r and ξ switched. We choose I_2 and begin by pulling $\eta_r(x)$ out in front of the integral. The $\eta_r(x)$ factor will be estimated using the L^∞ norm. Now we are left with a singular integral operator much like that of Theorem 5.12 except that the singular term is too singular. At this point we require the special form of K. In the case $K(x,y) = \frac{x-y}{|x-y|^{d+1}}$ we bring the factor $\frac{x-y}{|x-y|}$ out of the integral by using its L^∞ bound and thereby reduce to the case of $K(x,y) = \frac{1}{|x-y|^d}$. In this case we write

$$
\frac{K(x,y)}{|x-y|} = \frac{1}{|x-y|^{d+1}} = \operatorname{div}_y \left[\frac{x-y}{|x-y|^{d+1}} \right].
$$

We now integrate by parts which results in a kernel of the right singularity at the cost of derivatives appearing on the other terms in the integrand. Without loss of generality consider $K(x,y) = \frac{1}{|x-y|^d}$; then the integral I_2 becomes

$$
\begin{aligned}
I_2 &= \eta_r(x) \int \frac{K(x,y)}{|x-y|} \left(\prod_{j=1, j\neq r}^{p} q_1(\eta_j) \right) \xi(y)\, dy \\
&= -\eta_r(x) \left\{ \int \frac{(x-y)}{|x-y|^{d+1}} \cdot \nabla_y \left[\prod_{j=1, j\neq r}^{p} q_1(\eta_j) \right] \xi(y)\, dy \right. \\
&\qquad \left. + \int \left(\prod_{j=1, j\neq r}^{p} q_1(\eta_j) \right) \frac{(x-y)}{|x-y|^{d+1}} \cdot \nabla_y \left[\xi(y) \right]\, dy \right\} \\
&= -\eta_r(x) \left\{ \sum_{s=1, s\neq r}^{p} \int \frac{1}{|x-y|^d} \left(\prod_{j=1, j\neq r, s}^{p} q_1(\eta_j) \right) \frac{(x-y)}{|x-y|} \cdot \nabla_y \left[q_1(\eta_s) \right] \xi(y)\, dy \right. \\
&\qquad \left. + \int \frac{1}{|x-y|^d} \left(\prod_{j=1, j\neq r}^{p} q_1(\eta_j) \right) \frac{(x-y)}{|x-y|} \cdot \nabla_y \left[\xi(y) \right]\, dy \right\}.
\end{aligned}
$$

Now, from Lemmas 5.16 and 5.17 we can estimate these two terms as long as $\xi \in W^{1,q}$, $\eta_r \in L^\infty$, and $\eta_j \in C^2$ for $j \neq r$. As mentioned earlier, the term I_3 can be handled

analogously with the requirements that $\xi \in L^\infty$, $\eta_r \in W^{1,q}$, and $\eta_j \in C^2$ for $j \neq r$. The estimate in the statement of the theorem is now realized. ∎

As a precursor to estimating the operators $C_p(\eta)$ in $W^{s,q}$ norm, we estimate the following simplified operators:

$$S_p(\eta, \ldots, \eta)\ \xi(x) = \int k(x,y) q_1(\eta)^p \xi(y)\ dy.$$

These will lead to the operator $C_p(\eta)$ since the only difference is the presence in $C_p(\eta)$ of the analytic function $c_p(q_1(\eta))$ which we will expand in a Taylor series, giving rise to a sum of operators of the form $S_p(\eta, \ldots, \eta)$. With this in mind we prove the following theorem.

THEOREM 5.18. *If $d = 1$, the following estimate holds for $\eta \in W^{s+1,q}$ and $\xi \in W^{s,q}$, $s > \max(\frac{1}{q}, 2)$:*

$$\left\| \partial_x^s S_p(\eta, \ldots, \eta)\ \xi \right\|_{L^q} \leq C \left\| \eta \right\|_{W^{s+1,q}}^p \left\| \xi \right\|_{W^{s,q}}. \tag{5.23}$$

If $d = 2$, the following estimate holds for $\eta \in W^{s+2,q}$ and $\xi \in W^{s,q}$, $s > \max(\frac{2}{q}, 3)$:

$$\left\| \partial_x^s S_p(\eta, \ldots, \eta)\ \xi \right\|_{L^q} \leq C \left\| \eta \right\|_{W^{s+2,q}}^p \left\| \xi \right\|_{W^{s,q}}. \tag{5.24}$$

In order to prove this theorem we need the following lemma which is proven in W. Craig, U. Schanz, and C. Sulem [7] and D. Nicholls [19].

LEMMA 5.19. *Let S_p be defined as*

$$S_p(\eta, \ldots, \eta)\ \xi(x) = \int K(x,y) q_1(\eta)^p \xi(y)\ dy. \tag{5.25}$$

Then the lth derivative of this operator is

$$\partial_x^l S_p(\eta, \ldots, \eta)\xi(x) = \sum_{m=0}^{l} \sum_{\sum \alpha_k = m} S_p(\partial_x^{\alpha_1}\eta, \ldots, \partial_x^{\alpha_p}\eta)\ \partial_x^{l-m}\xi(x). \tag{5.26}$$

We are now ready to prove Theorem 5.18.

Proof of Theorem 5.18. By Lemma 5.19 we have

$$\left\| \partial_x^s S_p(\eta, \ldots, \eta)\xi(x) \right\|_{L^q} \leq \sum_{m=0}^{s} \sum_{\sum \alpha_k = m} \left\| S_p(\partial_x^{\alpha_1}\eta, \ldots, \partial_x^{\alpha_p}\eta)\ \partial_x^{s-m}\xi(x) \right\|_{L^q}.$$

Since we wish to avoid using a C^{s+1} derivative we diverge a little from the approach of W. Craig, U. Schanz, and C. Sulem [7] and split the previous sum into three parts. The first will contain all terms where at least one derivative hits the function ξ. The second will contain terms with no derivatives on ξ but not all s derivatives on one η. The final part will contain all terms where all s derivatives are on one of the η.

Each part will consist of terms of the form

$$\left\| S_p(\partial_x^{\alpha_1}\eta, \ldots, \partial_x^{\alpha_p}\eta)\ \partial_x^{s-m}\xi(x) \right\|_{L^q}.$$

If $m < s$, then at least one derivative hits ξ and we can use Theorem 5.12 to make the estimate

$$\begin{aligned} \left\| S_p(\partial_x^{\alpha_1}\eta, \ldots, \partial_x^{\alpha_p}\eta)\ \partial_x^{s-m}\xi(x) \right\|_{L^q} &\leq C \left(\prod_{j=1}^{p} |\partial_x^{\alpha_j}\eta|_{C^1} \right) \left\| \partial_x^{s-m}\xi \right\|_{L^q} \\ &\leq C\, |\eta|_{C^1}^{p-1}\, |\partial_x^m \eta|_{C^1} \left\| \partial_x^{s-m}\xi \right\|_{L^q}. \end{aligned}$$

The second line comes from the interpolation

$$|\partial_x^\alpha \eta|_{C^1} \leq C\,|\eta|_{C^1}^{1-\frac{\alpha}{m}}\,|\partial_x^m \eta|_{C^1}^{\frac{\alpha}{m}}\,.$$

By being wasteful with derivatives we can estimate this by $C\,|\eta|_{C^1}^{p-1}\,|\eta|_{C^s}\,\|\eta\|_{W^{s,q}}$. By our choice of s for $d=1$ we can estimate this by $C\,\|\eta\|_{W^{s+1,q}}^p\,\|\xi\|_{W^{s,q}}$. Similarly, by the choice of s for $d=2$ we can estimate this by $C\,\|\eta\|_{W^{s+2,q}}^p\,\|\xi\|_{W^{s,q}}$. If we consider $m=s$ but $\alpha_r \neq s$, then we make the estimate, based on Theorem 5.12

$$\begin{aligned}\|S_p(\partial_x^{\alpha_1}\eta,\ldots,\partial_x^{\alpha_p}\eta)\;\xi(x)\|_{L^q} &\leq C\left(\prod_{j=1}^p |\partial_x^{\alpha_j}\eta|_{C^1}\right)\|\xi\|_{L^q}\\ &\leq C\,|\eta|_{C^s}^p\,\|\xi\|_{L^q}\,.\end{aligned}$$

Again, by our choice of s we can make the appropriate estimate. Finally, in the case where $\alpha_r = s$ we use Theorem 5.13 to make the estimate

$$\begin{aligned}&\|S_p(\eta,\ldots,\partial_x^s\eta,\ldots,\eta)\;\xi(x)\|_{L^q}\\ &\quad\leq C\left(\prod_{j=1}^p |\partial_x^{\alpha_j}\eta|_{C^1}\right)\|\xi\|_{L^q}\\ &\quad\leq C\,|\eta|_{C^1}^{p-1}\,\{|\eta|_{C^1}\,|\xi|_{C^1}\,\|\partial_x^s\eta\|_{L^q}\\ &\qquad+|\eta|_{C^1}\left(|\partial_x^s\eta|_{L^\infty}\,\|\partial_x\xi\|_{L^q}+|\xi|_{L^\infty}\left\|\partial_x^{s-1}\eta\right\|_{L^q}\right)\\ &\qquad+(p-1)\,|\eta|_{C^2}\,(|\partial_x^s\eta|_{L^\infty}\,\|\partial_x\xi\|_{L^q}+|\xi|_{L^\infty}\,\|\partial_x^s\eta\|_{L^q})\}\,.\end{aligned}$$

Once again, without being careful about derivatives one can make the appropriate estimate given the choice of s. □

Finally, we are in a position to establish the analyticity of the singular integral operators C_p in one and two dimensions in the appropriate function spaces.

THEOREM 5.20. *If $d=1$, $\eta \in W^{s+1,q}$, and $s > \max(\frac{1}{q},2)$, then the singular integral operator $C_p(\eta)$ is bounded on $W^{s,q}$ and*

$$\|C_p(\eta)\;\xi\|_{W^{s,q}} \leq C\,\|\eta\|_{W^{s+1,q}}^p\,\|\xi\|_{W^{s,q}}\,. \tag{5.27}$$

Furthermore, the operator $C_p(\eta)$ is analytic as a mapping on $W^{s,q}$ and thus its Taylor series converges in operator norm. If $d=2$, $\eta \in W^{s+2,q}$, and $s > \max(\frac{2}{q},3)$, then the singular integral operator $C_p(\eta)$ is bounded on $W^{s,q}$ and

$$\|C_p(\eta)\;\xi\|_{W^{s,q}} \leq C\,\|\eta\|_{W^{s+2,q}}^p\,\|\xi\|_{W^{s,q}}\,. \tag{5.28}$$

Furthermore, the operator $C_p(\eta)$ is analytic as a mapping on $W^{s,q}$ and thus its Taylor series converges in operator norm.

Proof. The idea behind the proof is to expand the analytic function $c_p(z)$ in its Taylor series expansion to reduce C_p to an infinite sum of operators S_p which have the appropriate decay. In brief we write

$$\begin{aligned}C_p(\eta)\;\xi(x) &= \int k(x,y)c_p(q_1)\xi(y)\,dy\\ &= \sum_{l=p}^\infty \frac{c_p^{(l)}(0)}{l!}S_p(\eta,\ldots,\eta)\;\xi(x).\end{aligned}$$

The proof now proceeds in exactly the same manner as in W. Craig, U. Schanz, and C. Sulem [7] or D. Nicholls [19], where we use Theorem 5.18 rather than the $C^{s+1} - W^{s,q}$ estimates of these papers. ∎

For the smoothing integral operators $C_{p,h}$ there is a development which can be followed which is very similar to, though not as delicate as, the one presented above for the singular integral operators C_p. Due to the choices of s which we will make, the estimates which we need have already been established by W. Craig, U. Schanz, and C. Sulem in two dimensions [7], and in d dimensions by D. Nicholls [19]. We state the result for completeness, and present the corollary which we use.

THEOREM 5.21. *Let $p+\rho+\lambda > d$, and suppose that $|\eta|_{L^\infty} < hR_0$ and $|\eta|_{C^s} < \infty$. Then $C_{p,h}$ is bounded from L^q to $W^{s,q}$ and*

$$\|C_{p,h}(\eta)\ \xi\|_{W^{s,q}} < C\,|\eta|_{L^\infty}^{p-1}\,|\eta|_{C^s}\,\|\xi\|_{L^q}\,. \tag{5.29}$$

Furthermore, the operator $C_{p,h}$ is analytic as a mapping from L^q to $W^{s,q}$ in the set

$$\{\eta \mid |\eta|_{L^\infty} < hR_0 \text{ and } |\eta|_{C^s} < \infty\}. \tag{5.30}$$

Consequently, $C_{p,h}$ is represented by its Taylor series expansion.

The corollary that we use is the following.

COROLLARY 5.22. *If $d = 1$, $\eta \in W^{s+1,q}$, and $s > \max(\frac{1}{q}, 2)$, then the smoothing integral operator $C_{p,h}$ is bounded from L^q to $W^{s,q}$ and*

$$\|C_{p,h}(\eta)\ \xi\|_{W^{s,q}} < C\,\|\eta\|_{W^{s+1,q}}^{p}\,\|\xi\|_{L^q}\,. \tag{5.31}$$

Furthermore, the operator $C_{p,h}$ is analytic as a mapping from L^q to $W^{s,q}$ and thus is represented by its Taylor series expansion. If $d = 2$, $\eta \in W^{s+2,q}$, and $s > \max(\frac{2}{q}, 3)$, then the smoothing integral operator $C_{p,h}$ is bounded from L^q to $W^{s,q}$ and

$$\|C_{p,h}(\eta)\ \xi\|_{W^{s,q}} < C\,\|\eta\|_{W^{s+2,q}}^{p}\,\|\xi\|_{L^q}\,. \tag{5.32}$$

Furthermore, the operator $C_{p,h}$ is analytic as a mapping from L^q to $W^{s,q}$ and thus is represented by its Taylor series expansion.

Acknowledgment. The first author wishes to thank his wife Kristy for her love and support.

REFERENCES

[1] J.T. BEALE, *The existence of cnoidal water waves with surface tension*, J. Differential Equations, 31 (1979), pp. 230–263.
[2] H. BECKERT AND E. ZEIDLER, *Beiträge zur Theorie und Praxis freier Randwertaufgaben*, Akademie-Verlag, Berlin, 1971.
[3] M. CHRIST AND J. JOURNÉ, *Polynomial growth estimates for multilinear singular integral operators*, Acta Math., 159 (1987), pp. 51–80.
[4] R. COIFMAN AND Y. MEYER, *Nonlinear harmonic analysis and analytic dependence*, in AMS Proc. Symp. Pure Math., 43 (1985), pp. 71–78.
[5] W. CRAIG AND M. GROVES, *Hamiltonian long-wave scaling limits of the water-wave problem*, Wave Motion, 19 (1994), pp. 367–389.
[6] W. CRAIG AND D. HASKELL, *On the orbit space of the harmonic oscillator*, in preparation.
[7] W. CRAIG, U. SCHANZ, AND C. SULEM, *The modulational regime of three-dimensional water waves and the Davey-Stewartson system*, Ann. Inst. Henri Poincaré Anal. Non Linéaire, 14 (1997), pp. 615–667.
[8] W. CRAIG AND C. SULEM, *Numerical simulation of gravity waves*, J. Comput. Phys., 108 (1993), pp. 73–83.

[9] E. Fadell and P. Rabinowitz, *Generalized cohomological index theories for Lie group actions with an application to bifurcation questions for Hamiltonian systems*, Invent. Math., 45 (1978), pp. 139–174.
[10] M.D. Groves and A. Mielke, *A Spatial Dynamics Approach to Three-Dimensional Gravity-Capillary Steady Water Waves,* preprint, 1999.
[11] J. Hammack, N. Scheffner, and H. Segur, *Two dimensional periodic waves in shallow water*, J. Fluid Mech., 209 (1989), pp. 567–589.
[12] J. Hammack, D. McCallister, N. Scheffner, and H. Segur, *Two dimensional periodic waves in shallow water* II: *Asymmetric waves*, J. Fluid Mech., 285 (1995), pp. 95–122.
[13] M. Jones and J. Toland, *The bifurcation and secondary bifurcation of capillary gravity waves*, Proc. Royal Soc. London Ser. A, 399 (1985), pp. 391–417.
[14] M. Jones and J. Toland, *Symmetry and bifurcation of capillary gravity waves*, Arch. Rational Mech. Anal., 96 (1986), pp. 29–53.
[15] T. Levi-Civita, *Détermination rigoureuse des ondes permanentes d'ampleur finie*, Math. Ann., 93 (1925), pp. 264–314.
[16] P. Milewski and J.B. Keller, *Three dimensional water waves*, Stud. Appl. Math., 37 (1996), pp. 149–166.
[17] J. Moser, *Periodic orbits near an equilibrium and a theorem by Alan Weinstein*, Comm. Pure Appl. Math., 29 (1976), pp. 727–747.
[18] D. Nicholls, *Traveling water waves: Spectral continuation methods with parallel implementation*, J. Comput. Phys., 142 (1998), pp. 224–240.
[19] D. Nicholls, *Traveling Gravity Water Waves in Two and Three Dimensions*, Ph.D. Diss., Brown University, Providence, RI, 1998.
[20] D. Nicholls and W. Craig, *Traveling Gravity Water Waves in Two and Three Dimensions*, in preparation.
[21] P. Plotnikov, *Solvability of the problem of spatial gravitational waves on the surface of an ideal fluid*, Soviet Physics Dokl., 25 (1980), pp. 170–171.
[22] P. Rabinowitz, *Variational methods for nonlinear eigenvalue problems*, CIME, III ciclo, Varenna, 1974, Edizioni Cremonese, Rome, 1974, pp. 139–195.
[23] J. Reeder and M. Shinbrot, *Three dimensional nonlinear wave interaction in water of constant depth*, Nonlinear Anal., 5 (1981), pp. 303–323.
[24] J. Reeder and M. Shinbrot, *On Wilton ripples* II: *Rigorous results*, Arch. Rational Mech. Anal., 77 (1981), pp. 321–347.
[25] U. Schanz, *On the Evolution of Gravity-Capillary Waves in Three Dimensions.*, Diss., University of Toronto, Toronto, Ontario, 1997.
[26] J. Stoker, *Water Waves: The Mathematical Theory and Applications*, Interscience, New York, 1957.
[27] D. Struik, *Détermination rigoureuse des ondes irrotationnelles périodiques dans un canal à profondeur finie*, Math. Ann., 95 (1926), pp. 595–634.
[28] T.Y. Sun, *Three-dimensional steady water waves generated by partially localized pressure disturbances*, SIAM J. Math. Anal., 24 (1993), pp. 1153–1178.
[29] A. Weinstein, *Normal modes for nonlinear hamiltonian systems*, Invent. Math., 20 (1973), pp. 47–57.
[30] V.E. Zakharov, *Stability of periodic waves of finite amplitude on the surface of a deep fluid*, J. Appl. Mech. Tech. Phys., 9 (1968), pp. 190–194.
[31] E. Zeidler, *Existenzbeweis für cnoidal waves unter Berücksichtigung der Oberflächen spannung*, Arch. Rational Mech. Anal, 41 (1971), pp. 81–107.

SIAM J. MATH. ANAL.
Vol. 32, No. 2, pp. 360–384

STABILITY AND HOPF BIFURCATION FOR FULLY NONLINEAR PARABOLIC-HYPERBOLIC EQUATIONS*

HERBERT KOCH† AND STUART S. ANTMAN‡

Abstract. We develop the principle of linearized stability and a Hopf bifurcation theorem as elements of a geometric theory for fully nonlinear parabolic-hyperbolic problems. Crucial steps in our work are showing the differentiability of the time-t map, showing that the admissible initial data form a manifold (whose failure to be linear is due to the general boundary conditions we study), and analyzing the spectrum of the generator of the linearized semigroup. This paper provides the abstract framework for the study of a class of concrete problems of self-sustained oscillations of nonlinearly viscoelastic bodies like that treated by Antman and Koch [*SIAM J. Appl. Math.*, 60 (2000), pp. 1357–1387]. Our equations are intrinsically interesting: They provide an example of a new kind of semiflow that combines properties of ordinary differential equations and parabolic equations in a novel way.

Key words. Hopf bifurcation, linearized stability, fully nonlinear parabolic-hyperbolic equations, oblique boundary conditions, nonlinear viscoelasticity

AMS subject classifications. 35B10, 35B32, 35B35, 35K55, 35K70, 47H20, 58G28, 73G25

PII. S003614109833793X

1. Introduction.

In this paper we extend the so-called geometrical theory of parabolic equations of Henry [13] to fully nonlinear parabolic-hyperbolic equations with fully nonlinear boundary conditions, which include oblique boundary conditions. We establish well-posedness and determine the structure of the local semiflow near an equilibrium. A main motivation for the theory we develop is to treat self-sustained oscillations of nonlinearly viscoelastic solids (cf. [7]). A feature of such problems, which is illustrated in our examples, is that the boundary conditions have a very rich structure.

Notation. Lower-case boldface symbols represent n-tuples of real numbers, e.g., $\mathbf{x} = (x_1, \dots, x_n)$, or functions with values in $\mathbb{R}^n$. Upper-case boldface symbols represent $n \times n$ matrices, e.g., $\mathbf{Q} = (q_{ij})$, or functions with such values. We denote the $\frac{1}{2}n(n+1)$-dimensional space of symmetric $n \times n$ matrices by Sym^n. The zero and identity matrices are denoted $\mathbf{O}$ and $\mathbf{I}$. We denote the inner product of two n-tuples $\mathbf{a}$ and $\mathbf{b}$ by $\mathbf{a} \cdot \mathbf{b} \equiv \sum_{j=1}^n a_j b_j$ and denote the inner product of two $n \times n$ matrices $\mathbf{A}$ and $\mathbf{B}$ by $\mathbf{A} : \mathbf{B} \equiv \sum_{j,k=1}^n A_{jk} B_{jk}$. We also set $\mathbf{A} : \mathbf{ab} \equiv \mathbf{a} \cdot \mathbf{A} \cdot \mathbf{b} \equiv \sum_{j,k=1}^n A_{jk} a_j b_k$. Any matrix $\mathbf{A}$, symmetric or not, is said to be *positive-definite* if its quadratic form $\mathbf{a} \cdot \mathbf{A} \cdot \mathbf{a}$ (which only involves the symmetric part of $\mathbf{A}$) is positive-definite. We denote the (infrequently used) zero third-order tensor by $\mathbf{O}$.

The gradient $(\partial u/\partial x_1, \dots, \partial u/\partial x_n)$ of a scalar-valued function u of $\mathbf{x}$ is denoted by $u_{\mathbf{x}}$. Likewise, the matrix $\left(\partial^2 u/\partial x_i \partial x_j\right)$ of second partial derivatives of u with respect to $\mathbf{x}$ is denoted by $u_{\mathbf{xx}}$. The gradient of a scalar-valued function F with

*Received by the editors April 29, 1998; accepted for publication (in revised form) September 23, 1999; published electronically July 5, 2000.

http://www.siam.org/journals/sima/32-2/33793.html

†Institut für Angewandte Mathematik, Universität Heidelberg, D-69120 Heidelberg, Germany (koch@iwr.uni-heidelberg.de). The work of this author was supported in part by the DFG.

‡Department of Mathematics and Institute for Physical Science and Technology, University of Maryland, College Park, MD 20742-4015 (ssa@math.umd.edu). The work of this author was supported in part by the NSF and by ARO-MURI97 Grant DAAG55-97-1-0114 to the Center for Dynamics and Control of Smart Structures.

respect to a symmetric matrix argument $\mathbf{Q}$ is the symmetric matrix denoted by $F_{\mathbf{Q}}$. It is defined to be the unique symmetric matrix $\mathbf{A}$ such that $\partial_\varepsilon F(\mathbf{Q}+\varepsilon\mathbf{B})|_{\varepsilon=0} = \mathbf{A}:\mathbf{B}$ for all symmetric $\mathbf{B}$. We employ obvious generalizations of these conventions.

Let $\mathcal{A}$ and $\mathcal{B}$ be open subsets of Banach spaces, let $k \geq 0$ be an integer, and let $\sigma \in (0,1]$. Then $C^k(\mathcal{A},\mathcal{B})$ denotes the space of all k-times continuously differentiable functions from $\mathcal{A}$ to $\mathcal{B}$. If $\mathcal{A}$ and $\mathcal{B}$ are subsets of $\mathbb{R}^n$, then $C^{k,\sigma}(\mathcal{A},\mathcal{B})$ denotes the space of all k-times continuously differentiable functions from $\mathcal{A}$ to $\mathcal{B}$ whose kth order derivatives are Hölder continuous with exponent σ. When $\mathcal{B}$ is obvious, we suppress it. We extend this notation in the obvious way if $\mathcal{A}$ is contained in the closure of its interior.

We now formulate our initial-boundary-value problem. Let $\Omega \subset \mathbb{R}^n$ be a bounded domain with smooth boundary Γ and let T be a positive number. We set $\Omega_T = \Omega \times [0,T)$ and $\Gamma_T = \Gamma \times [0,T)$. We shall study an initial-boundary-value problem for a scalar-valued function $\Omega \times [0,T] \ni (\mathbf{x},t) \mapsto u(\mathbf{x},t)$. Let $\mathcal{I}$ be an interval containing 0 and let ζ be a parameter in $\mathcal{I}$. The domains of the functions defining the differential equation and the boundary condition for our problem are

$$\mathcal{F} = \bar{\Omega} \times (\mathbb{R} \times \mathbb{R}^n \times \mathrm{Sym}^n)^2 \times \mathcal{I}, \quad \mathcal{B} = \Gamma \times (\mathbb{R} \times \mathbb{R}^n)^2 \times \mathcal{I}. \tag{1.1}$$

We are given the functions

$$\mathcal{F} \ni (\mathbf{x},u,\mathbf{p},\mathbf{Q},\dot{u},\dot{\mathbf{p}},\dot{\mathbf{Q}},\zeta) \mapsto F(\mathbf{x},u,\mathbf{p},\mathbf{Q},\dot{u},\dot{\mathbf{p}},\dot{\mathbf{Q}},\zeta) \in C^{k+1}(\mathcal{F},\mathbb{R}), \tag{1.2a}$$

$$\mathcal{B} \ni (\mathbf{x},u,\mathbf{p},\dot{u},\dot{\mathbf{p}},\zeta) \mapsto B(\mathbf{x},u,\mathbf{p},\dot{u},\dot{\mathbf{p}},\zeta) \in C^{k+2}(\mathcal{B},\mathbb{R}). \tag{1.2b}$$

In (1.2) the dots appearing over u, $\mathbf{p}$, $\mathbf{Q}$ have no operational significance; the symbols $\dot{u}$, $\dot{\mathbf{p}}$, $\dot{\mathbf{Q}}$ merely identify the arguments to be filled by functions that are the derivatives of the functions filling the slots occupied by u, $\mathbf{p}$, $\mathbf{Q}$.

We study the initial-boundary-value problem

$$u_{tt} - F(\mathbf{x},u,u_{\mathbf{x}},u_{\mathbf{xx}},u_t,u_{\mathbf{x}t},u_{\mathbf{xx}t},\zeta) = 0 \quad \text{in } \Omega_T, \tag{1.3a}$$

$$B(\mathbf{x},u,u_{\mathbf{x}},u_t,u_{\mathbf{x}t},\zeta) = 0 \quad \text{on } \Gamma_T, \tag{1.3b}$$

$$u(\cdot,0) = u_0(\cdot) \quad \text{in } \Omega, \tag{1.3c}$$

$$u_t(\cdot,0) = u_1(\cdot) \quad \text{in } \Omega. \tag{1.3d}$$

Note that (1.3b) can be specialized to a Dirichlet condition in which u is prescribed, or to a Neumann condition in which the normal derivative of u is prescribed, or to a mixed condition in which the normal derivative of u is a prescribed function of u. For the applications we want to treat, we need much of the generality of (1.3b), which we may term an *oblique* boundary condition because it involves derivatives other than normal derivatives to Γ_T in space-time. Of course, the nature of the boundary condition can vary from point to point on the boundary, but our smoothness assumptions in (1.2b) prevent, e.g., a Dirichlet condition being specified on a region of Γ_T that touches another such region on which a Neumann condition is specified. In our applications in section 2, Γ is a union of disjoint hypersurfaces in $\mathbb{R}^n$. For such Γ's we can prescribe different kinds of boundary conditions on each component.

For interpreting semiflows and semigroups, it is convenient to replace (1.3) with an equivalent system for the pair $(u,\dot{u})$ containing only first derivatives with respect to t:

$$
\begin{aligned}
u_t - \dot{u} &= 0 && \text{in } \Omega_T,\\
\dot{u}_t - F(\mathbf{x}, u, u_{\mathbf{x}}, u_{\mathbf{xx}}, \dot{u}, \dot{u}_{\mathbf{x}}, \dot{u}_{\mathbf{xx}}, \zeta) &= 0 && \text{in } \Omega_T,\\
B(\mathbf{x}, u, u_{\mathbf{x}}, \dot{u}, \dot{u}_{\mathbf{x}}, \zeta) &= 0 && \text{on } \Gamma_T,\\
u(\cdot, 0) &= u_0(\cdot) && \text{in } \Omega,\\
\dot{u}(\cdot, 0) &= u_1(\cdot) && \text{in } \Omega.
\end{aligned} \tag{1.4}
$$

Our basic assumption for ensuring that (1.3) or (1.4) be dissipative is the following.

HYPOTHESIS 1.1. *The symmetric matrix $F_{\dot{\mathbf{Q}}}$ is positive-definite. The vector $B_{\dot{\mathbf{p}}}$ is transversal to Γ.*

The basic result that enables us to talk about solutions to this problem and to formulate conditions for their periodicity is the following.

THEOREM 1.2. *Let Hypothesis* 1.1 *hold and let $u_0, u_1 \in C^{2,\sigma}(\bar{\Omega})$ satisfy the compatibility condition*

$$
B(\mathbf{x}, u_0, u_{0\mathbf{x}}, u_1, u_{1\mathbf{x}}) = 0 \quad \textit{for} \quad \mathbf{x} \in \Gamma. \tag{1.5}
$$

Then for T sufficiently small, there exists a unique solution to (1.3) *for which u_{tt} and $u_{\mathbf{x}\mathbf{x}t}$ are Hölder continuous.*

The corresponding solution $(u, \dot{u})$ of (1.4) is denoted

$$
(u(\cdot, t), \dot{u}(\cdot, t)) = \boldsymbol{\phi}(t, u_0, u_1, \zeta). \tag{1.6}
$$

In section 2, we give some examples of quasi-linear equations of the form (1.3) from solid mechanics. We prove a linearized stability principle (Theorem 3.6) and a Hopf bifurcation theorem (Theorem 3.8) under further assumptions on the dependence on the parameter ζ of solutions of the linearization of (1.3) about a steady solution, but we postpone to section 3 a detailed statement of the results because they depend on the introduction of further technical concepts. Section 3, the last of the three introductory sections, gives precise statements of these and other theorems, and organizes the logical structure of the paper.

During the last two decades, there has been an extensive development of the geometrical theory for partial differential equations. We mention the theory for weak solutions for quasilinear parabolic equations of Amann and the work on parabolic equations of Lunardi. See Amann [2], [3], [4] and Lunardi [19], which contain extensive references.

The theory of parabolic-hyperbolic equations can be subsumed under the theory of analytic semigroups for nonlinear equations. Nevertheless, the semiflows for parabolic-hyperbolic equations are strikingly different from those for parabolic equations. In particular, local semiflows for parabolic-hyperbolic equations do not regularize, although the autonomous linearized problem still defines an analytic semigroup.

We treat initial data that merely form a Banach manifold, rather than a Banach space, in order to accommodate general boundary conditions of the sort that arise in the concrete applications described in section 2. It consequently seems that neither our problem nor corresponding problems for purely parabolic equations can be treated by the available abstract methods for the latter, because these methods cannot handle general boundary conditions (1.3b) under natural structure conditions. We derive geometric properties of the local semiflow by first establishing differentiability of the semiflow.

The semiflow is not compact because the spectrum of the generator of the semigroup for linear equations contains points other than eigenvalues of finite multiplicity.

The control of this additional spectrum is very delicate. Even for scalar equations, we need the deep results of Agmon, Douglis, and Nirenberg [1] to clarify the behavior of the essential spectrum.

We also need a new proof of the Hopf bifurcation theorem that is closer to the original proof of Hopf [14] (see Howard and Kopell [15]) than to those used by Crandall and Rabinowitz [9] or Guckenheimer and Holmes [12]: We characterize solutions of period T by

$$\boldsymbol{\phi}(T, u, \dot{u}, \zeta) = (u, \dot{u}). \tag{1.7}$$

We use bifurcation methods to solve this equation. Obviously this approach only requires regularity of the semiflow $\boldsymbol{\phi}$ together with the usual conditions on the spectrum (but now formulated for the linearized semigroup).

Our major concerns are showing that the admissible initial data form a Banach manifold, which is generally not a Banach space, demonstrating the regularity of the flow, and controlling the spectrum of the semigroup generated by linear equations. In confronting these issues, we strive to make the exposition as flexible as possible.

We restrict our attention to scalar equations. It is, however, clear that the same arguments apply as well for general systems with mixed Dirichlet–Neumann boundary conditions. The requisite modifications will be discussed where they are necessary. The optimal smoothness of the constitutive functions F and B is of minor importance. Only in Theorem 3.3 do we bother to impose smoothness assumptions that require a treatment different from that for C^∞ constitutive functions.

For simplicity, we consider only strong solutions in the Hölder spaces $C^{2,\sigma}$, neglecting the available divergence structure in the physical models. The arguments can easily be adapted to weak solutions. The crucial results of Agmon, Douglis, and Nirenberg [1] and Solonnikov [24], however, are not available in that generality for weak solutions, though similar estimates for weak solutions no doubt hold under appropriate conditions.

2. Examples. We now give some examples of quasilinear problems for nonlinearly viscoelastic bodies, to which our theory can readily be applied.

2.1. Shearing of a viscoelastic layer due to friction on its face. Let $u(x,t)$ denote the displacement transverse to the x-axis of a layer of an incompressible viscoelastic material of strain-rate type. Then u satisfies

$$\begin{aligned} &\rho(x)u_{tt} = \sigma(u_x, u_{xt}, x)_x, \\ u(0,t) = 0, \qquad &\sigma(u_x(1,t), u_{xt}(1,t), 1) = f(\zeta - u_t(1,t)), \end{aligned} \tag{2.1}$$

where ρ is the density, where σ is the given constitutive function delivering the shear stress with $\sigma_b(a,b,x) \geq \text{const} > 0$ for all a, b, x, and where $-f(v)$ is the given friction force produced by a belt moving on the boundary $x = 1$ with velocity v relative to that of the viscoelastic layer. Here ζ is the actual speed of the belt. This problem for f of stick-slip type is treated in [7].

2.2. Antiplane shearing of a viscoelastic tube due to friction on one bounding surface. A version of (2.1) with two independent spatial variables $(x_1, x_2) = \mathbf{x}$ corresponds to the antiplane motion of an infinite viscoelastic tube with a (doubly-

connected) cross section Ω lying between bounding curves $\mathcal{S}_1$ and $\mathcal{S}_2$:

$$\begin{aligned} \rho(\mathbf{x})u_{tt} &= \frac{\partial}{\partial x_1}\sigma_1(u_{\mathbf{x}}, u_{\mathbf{x}t}, \mathbf{x}) + \frac{\partial}{\partial x_2}\sigma_2(u_{\mathbf{x}}, u_{\mathbf{x}t}, \mathbf{x}), \\ u(\mathbf{x}, t) &= 0 \quad \text{for} \quad \mathbf{x} \in \mathcal{S}_1, \\ \nu_1\sigma_1(u_{\mathbf{x}}, u_{\mathbf{x}t}, \mathbf{x}) + \nu_2\sigma_2(u_{\mathbf{x}}, u_{\mathbf{x}t}, \mathbf{x}) &= f(\zeta - u_t(\mathbf{x}, t), \mathbf{x}) \quad \text{for} \quad \mathbf{x} \in \mathcal{S}_2, \end{aligned} \tag{2.2}$$

where σ_1 and σ_2 are given constitutive functions satisfying Hypothesis 1.1, and where (ν_1, ν_2) is the unit outer normal to $\mathcal{S}_2$.

Problems in nonlinear solid mechanics, like these, in which there is but one dependent variable are rare. Typically, there are several dependent variables, as in the version of problem (2.1) for compressible materials and as in the following example.

2.3. Beck's problem for an extensible, shearable, nonlinearly viscoelastic rod. We study the deformation in a plane of a naturally straight rod of scaled unit length 1. Let $\boldsymbol{i}, \boldsymbol{j}$ be an orthonormal pair of vectors. In our model, a configuration of the rod at time t is specified by a curve $[0,1] \ni x \mapsto \boldsymbol{r}(x,t) \in \operatorname{span}\{\boldsymbol{i}, \boldsymbol{j}\}$ and a unit-vector field $[0,1] \ni x \mapsto \boldsymbol{a}(\theta(x,t)) \equiv \cos\theta(x,t)\,\boldsymbol{i} + \sin\theta(x,t)\,\boldsymbol{j}$. The vector $\boldsymbol{r}(x,t)$ may be interpreted as the position at time t of a material point lying along the line of centroids of the rod in its straight, natural configuration at a distance x from one end. The vector $\boldsymbol{a}(\theta(x,t))$ may be interpreted as the unit normal at time t to the deformed image of the section at x. We set $\boldsymbol{b}(\theta(x,t)) = -\sin\theta(x,t)\boldsymbol{i} + \cos\theta(x,t)\,\boldsymbol{j}$. If the rod is viscoelastic of strain-rate type and if the only loads on it are applied at its ends, then its geometrically exact equations of motion have the form

$$\begin{aligned} A(x)\boldsymbol{r}_{tt} &= [N(\nu, \eta, \theta_x, \nu_t, \eta_t, \theta_{xt}, x)\boldsymbol{a}(\theta) + H(\nu, \eta, \theta_x, \nu_t, \eta_t, \theta_{xt}, x)\boldsymbol{b}(\theta)]_x, \\ I(x)\theta_{tt} &= [M(\nu, \eta, \theta_x, \nu_t, \eta_t, \theta_{xt}, x)]_x \\ &\quad + \nu H(\nu, \eta, \theta_x, \nu_t, \eta_t, \theta_{xt}, x) - \eta N(\nu, \eta, \theta_x, \nu_t, \eta_t, \theta_{xt}, x), \end{aligned} \tag{2.3}$$

where

$$\nu \equiv \boldsymbol{r}_x \cdot \boldsymbol{a}, \qquad \eta \equiv \boldsymbol{r}_x \cdot \boldsymbol{b},$$

where A and I are given positive-valued functions and where

$$(\nu, \eta, \mu, \dot\nu, \dot\eta, \dot\mu, x) \mapsto N(\nu, \eta, \mu, \dot\nu, \dot\eta, \dot\mu, x), H(\nu, \eta, \mu, \dot\nu, \dot\eta, \dot\mu, x), M(\nu, \eta, \mu, \dot\nu, \dot\eta, \dot\mu, x)$$

are given constitutive functions with (the symmetric part of)

$$\begin{pmatrix} N_{\dot\nu} & N_{\dot\eta} & N_{\dot\mu} \\ H_{\dot\nu} & H_{\dot\eta} & H_{\dot\mu} \\ M_{\dot\nu} & M_{\dot\eta} & M_{\dot\mu} \end{pmatrix}$$

uniformly positive-definite. When their arguments are evaluated at (x,t), $H\boldsymbol{a} + H\boldsymbol{b}$ is the contact force and M is the contact couple exerted at time t by the material to the right of x on the material to the left of x. We assume that the end $x = 0$ of the rod is welded to the normal to $\boldsymbol{i}$ at $\mathbf{0}$, so that

$$\boldsymbol{r}(0,t) = \mathbf{0}, \qquad \theta(0,t) = 0. \tag{2.4}$$

We assume that the end $x = 1$ is hinged and is subject to a compressive follower load of magnitude ζ which always acts normal to the section at $x = 1$, so that

$$\begin{aligned} N(\nu(1,t), \eta(1,t), \theta_x(1,t), \nu_t(1,t), \eta_t(1,t), \theta_{xt}(1,t), 1) &= -\zeta, \\ H(\nu(1,t), \eta(1,t), \theta_x(1,t), \nu_t(1,t), \eta_t(1,t), \theta_{xt}(1,t), 1) &= 0, \\ M(\nu(1,t), \eta(1,t), \theta_x(1,t), \nu_t(1,t), \eta_t(1,t), \theta_{xt}(1,t), 1) &= 0. \end{aligned} \tag{2.5}$$

It is this interesting boundary condition and variants thereof that characterize Beck's problem. (For a discussion of the formulation of this problem, its generalizations to deformations in space, and some of the elementary properties of its solutions, see Antman [5] and Antman and Kenney [6].)

3. Outline and main results. We now give a detailed account of our results and methods.

Let $\mathcal{Z}$ be a topological space of parameters ζ. Let $\mathcal{X}$ be a Hausdorff space. For each $\zeta \in \mathcal{Z}$, let $\mathcal{U}(\zeta)$ be a subset of $\mathcal{X}$, whose elements $\mathbf{z}$ we shall identify with the pair (u_0, u_1). A *parametrized local semiflow* on $\mathcal{U} = \{(\mathbf{z}, \zeta) : \mathbf{z} \in \mathcal{U}(\zeta), \zeta \in \mathcal{Z}\}$ is a function pair $(\boldsymbol{\phi}, \hat{T})$, where $\hat{T}(\cdot, \zeta)$ is a lower-semicontinuous positive-valued function from $\mathcal{U}(\zeta)$ to the extended real line, with

$$\begin{gathered}\boldsymbol{\phi} \in C(\{(t, \mathbf{z}) : \mathbf{z} \in \mathcal{U},\ 0 \le t < \hat{T}(\mathbf{z}, \zeta)\},\ \mathcal{U}(\zeta)),\\ \hat{T}(\boldsymbol{\phi}(t, \mathbf{z}, \zeta)) + t = \hat{T}(\mathbf{z}, \zeta) \quad \text{if} \quad 0 \le t < \hat{T}(\mathbf{z}, \zeta),\\ \boldsymbol{\phi}(0, \mathbf{z}, \zeta) = \mathbf{z},\\ \boldsymbol{\phi}(t + s, \mathbf{z}, \zeta) = \boldsymbol{\phi}(s, \boldsymbol{\phi}(t, \mathbf{z}, \zeta), \zeta) \quad \text{if} \quad s + t < \hat{T}(\mathbf{z}, \zeta).\end{gathered}$$

$\hat{T}(\mathbf{z}, \zeta)$ is the *life span* of the trajectory $\boldsymbol{\phi}(\cdot, \mathbf{z}, \zeta)$ starting at $\mathbf{z}$ at time $t = 0$.

For $\sigma \in (0, 1]$, let $h^{k,\sigma}$ be the closure of C^∞ in $C^{k,\sigma}$ and let

$$\mathcal{H}^{2,\sigma}(\zeta) = \{(u_0, u_1) \in (h^{2,\sigma})^2 : B(\mathbf{x}, u_0, u_{0\mathbf{x}}, u_1, u_{1\mathbf{x}}, \zeta) = 0 \text{ if } \mathbf{x} \in \Gamma\}. \tag{3.1}$$

Let us identify $\mathbf{z} = (u_0, u_1)$, $\mathcal{U}(\zeta) = \mathcal{H}^{2,\sigma}(\zeta)$, and $\mathcal{Z} = \mathcal{I}$ where $\mathcal{I}$ is an interval of $\mathbb{R}$ containing 0. We shall prove the following more precise variant of Theorem 1.2 below.

THEOREM 3.1. *Problem* (1.4) *defines a parametrized local semiflow* $\boldsymbol{\phi}$ *on* $\mathcal{U}$.

A crucial ingredient for a geometric theory is a differentiable structure on $\mathcal{H}^{2,\sigma}(\zeta)$ and on $\mathcal{U}$. We say that a subset $\mathcal{E}$ of a Banach space $\mathcal{X}$ is a *C^k-Banach submanifold* of $\mathcal{X}$ if there is a Banach space $\mathcal{Y}$, a C^k map $\mathcal{X} \ni \mathbf{w} \mapsto \psi(\mathbf{w}) \in \mathcal{Y}$ with its derivative $\psi_{\mathbf{w}}$ surjective, and a linear injective map $L : \mathcal{Y} \to \mathcal{X}$ such that

$$\mathcal{E} = \psi^{-1}(\{0\}), \quad \mathcal{X} = [\psi_{\mathbf{w}}(\mathbf{w})]^{-1}(\{0\}) \oplus L(\mathcal{Y}) \quad \text{for all } \mathbf{w} \in \mathcal{E}. \tag{3.2}$$

In general, one should localize this notion, and for global considerations, one should require paracompactness, weak continuity, or both. This global naive definition however suffices for our purposes because we want to carry out a local analysis by using the chain rule, the implicit-function theorem, and the contraction-mapping principle. It is a simple exercise to show that classical constructions for finite-dimensional manifolds provide local structure maps in our setting. The map L always exists if $\mathcal{X}$ is finite-dimensional. At the level considered in this paper, the main difference between finite-dimensional and infinite-dimensional subspaces consists in the necessity of requiring the existence of L ab initio for the latter: Submanifolds certainly do not have better properties than linear subspaces, which do not necessarily have complementing subspaces.

THEOREM 3.2. *The set* $\mathcal{U} = \{(\mathbf{z}, \zeta) : \mathbf{z} \in \mathcal{H}^{2,\sigma}(\zeta),\ \zeta \in \mathcal{Z}\}$ *is a* C^k *Banach submanifold of* $(h^{2,\sigma})^2 \times \mathcal{I}$ *(in consequence of the choices* $\mathcal{Y} = h^{1,\sigma}(\Gamma)$, $\psi(u, v; \zeta) := B(\mathbf{x}, u, u_{\mathbf{x}}, v, v_{\mathbf{x}}; \zeta)$, *and* $Lg = (0, w, 0)$, *where* $w|_\Gamma = 0$ *and* $\boldsymbol{\nu} \cdot w_{\mathbf{x}} = g$ *on* Γ*). Moreover,* $\mathcal{H}^{2,\sigma}(\zeta)$ *is a* C^k*-Banach submanifold of* $h^{2,\sigma}$ *for all* ζ *with the same space* $\mathcal{Y}$ *and the same operator* L *for all* ζ *and structure maps depending continuously on* ζ, *i.e., the map* $\mathcal{I} \ni \zeta \mapsto \psi(\cdot, \zeta) \in C^1$ *is continuous.*

We now describe the regularity of the semiflow. On Ω_T we define Hölder spaces $C_{\mathrm{P}}^{j,\sigma}(\Omega_T)$, $j = 0, 1, 2$, for parabolic equations by the norms

(3.3a)
$$\|u\|_{C_{\mathrm{P}}^{0,\sigma}(\Omega_T)} = \max\left\{\|u\|_\infty, \sup_{0<|\mathbf{x}-\mathbf{y}|^2+|t-s|\le 1}\left\{\frac{|u(\mathbf{x},t)-u(\mathbf{y},s)|}{(|\mathbf{x}-\mathbf{y}|^2+|t-s|)^{\sigma/2}}\right\}\right\},$$

(3.3b)
$$\|u\|_{C_{\mathrm{P}}^{1,\sigma}(\Omega_T)} = \max\left\{\|u\|_\infty, \|u_{\mathbf{x}}\|_{C_{\mathrm{P}}^{0,\sigma}(\Omega_T)}, \sup_{\mathbf{x},\,0<|t-s|\le 1}\left\{\frac{|u(\mathbf{x},t)-u(\mathbf{x},s)|}{|t-s|^{(1+\sigma)/2}}\right\}\right\},$$

(3.3c)
$$\|u\|_{C_{\mathrm{P}}^{2,\sigma}(\Omega_T)} = \max\left\{\|u\|_\infty, \|u_t\|_{C_{\mathrm{P}}^{0,\sigma}(\Omega_T)}, \|u_{\mathbf{x}}\|_{C_{\mathrm{P}}^{1,\sigma}(\Omega_T)}\right\}.$$

Denote the closures of C^∞ in $C_{\mathrm{P}}^{j,\sigma}$ by $h_{\mathrm{P}}^{j,\sigma}$. We set

$$Y^{2,\sigma}(\Omega_T) = \{u : u, u_t \in h_{\mathrm{P}}^{2,\sigma}(\Omega_T)\} \tag{3.4a}$$

with the norm

$$\|u\|_{Y^{2,\sigma}} = \max\{\|u\|_{C_{\mathrm{P}}^{2,\sigma}}, \|u_t\|_{C_{\mathrm{P}}^{2,\sigma}}\}. \tag{3.4b}$$

THEOREM 3.3. *The local semiflow $\boldsymbol{\phi}$ for* (1.3) *is of class C^k, i.e., for $(u_0, u_1) \in \mathcal{H}^{2,\sigma}(\zeta)$, there is a neighborhood $\mathcal{V}$ of (u_0, u_1, ζ) in $\mathcal{U}$ and a $t_0 > 0$ such that*

$$[(v_0, v_1, \zeta) \mapsto \boldsymbol{\phi}(\cdot, v_0, v_1, \zeta)] \in C^k(\mathcal{V}, Y^{2,\sigma}(\Omega_{t_0})), \tag{3.5a}$$
$$\boldsymbol{\phi}(\cdot,\cdot,\cdot,\zeta)|_{(0,t_0)\times\mathcal{V}} \in C^k(\mathcal{U}\times(0,t_0), (C^{2,\sigma}(\Omega))^2). \tag{3.5b}$$

We continue with the study of the semiflow near a fixed point. Suppose that $u^0(\cdot,\zeta) \in C^3(\bar\Omega)$ is a stationary solution of (1.3). The linearization of (1.3) about u^0 is

$$v_{tt} - \mathbf{G} : v_{\mathbf{xx}} - \mathbf{g}\cdot v_{\mathbf{x}} - \gamma v - \mathbf{H} : v_{\mathbf{x}\mathbf{x}t} - \mathbf{h}\cdot v_{\mathbf{x}t} - \eta v_t = 0 \quad \text{in } \Omega_T, \tag{3.6a}$$
$$\mathbf{a}\cdot v_{\mathbf{x}} + \alpha v + \mathbf{b}\cdot v_{\mathbf{x}t} + \beta v_t = 0 \quad \text{on } \Gamma_T, \tag{3.6b}$$
$$v(\cdot,0) = v_0(\cdot) \quad \text{in } \Omega, \tag{3.6c}$$
$$v_t(\cdot,0) = v_1(\cdot) \quad \text{in } \Omega, \tag{3.6d}$$

where

$$\begin{array}{lll} \mathbf{G} \equiv F_{\mathbf{P}}, & \mathbf{g} \equiv F_{\mathbf{p}}, & \gamma \equiv F_u, \\ \mathbf{H} \equiv F_{\dot{\mathbf{P}}}, & \mathbf{h} \equiv F_{\dot{\mathbf{p}}}, & \eta \equiv F_{\dot u}, \\ & \mathbf{a} \equiv B_{\mathbf{p}}, & \alpha \equiv B_u, \\ & \mathbf{b} \equiv B_{\dot{\mathbf{p}}}, & \beta \equiv B_{\dot u}, \end{array} \tag{3.6e}$$

where the arguments of F and its derivatives are

$$(\mathbf{x}, u^0(\mathbf{x},\zeta), u^0_{\mathbf{x}}(\mathbf{x},\zeta), u^0_{\mathbf{xx}}(\mathbf{x},\zeta), 0, \mathbf{0}, \mathbf{O}, \zeta)$$

and those of B and its derivatives are

$$(\mathbf{x}, u^0(\mathbf{x},\zeta), u^0_{\mathbf{x}}(\mathbf{x},\zeta), 0, \mathbf{0}, \zeta).$$

Just as in (1.3), we can replace this problem with one containing just first derivatives with respect to t. We make this replacement without comment.

We define

$$\mathcal{H}_0^{2,\sigma}(\zeta) = \left\{(v_0, v_1) \in (h^{2,\sigma})^2 : \ \mathbf{a} \cdot v_{0\mathbf{x}} + \alpha v_0 + \mathbf{b} \cdot \ v_{1\mathbf{x}} + \beta v_1 = 0 \quad \text{on} \quad \Gamma\right\}, \tag{3.7}$$

where the arguments of $\mathbf{a}, \alpha, \mathbf{b}, \beta$ are those just shown. Then we have the following.

THEOREM 3.4. *Problem* (3.6) *defines an analytic semigroup* $t \mapsto S(t;\zeta)$ *on* $\mathcal{H}_0^{2,\sigma}(\zeta)$.

Let $\Sigma(L)$ denote the spectrum of a linear operator L.

COROLLARY 3.5. *Denote the generator of* $S(\cdot,\zeta)$ *by* $A(\zeta)$. *Then*

$$\Sigma(S(t,\zeta)) = e^{\Sigma(tA(\zeta))} \cup \{0\}.$$

Remark. There is a small imprecision in our notation: In contrast to the situation for semigroups for parabolic equations, the spectrum of $A(\zeta)$ may depend on the space (as may happen for hyperbolic equations; see [8]). We shall see that this is not the case under the reasonable assumptions we shall impose, thereby justifying a posteriori the notation of the corollary. The corollary is an immediate consequence of analyticity, up to the statement that 0 is in the spectrum of the semigroup. This will be a consequence of Proposition 4.5 below.

THEOREM 3.6 (Principle of linearized stability). *The following statements are equivalent:*

(i) $\Sigma(A(\zeta)) \subset \{z \in \mathbb{C} : \Re\, z < 0\}$.

(ii) *The trivial solution of* (3.6) *is exponentially stable.*

(iii) *The stationary solution* u^0 *of* (1.3) *is exponentially stable, i.e., solutions of* (1.3) *starting sufficiently close to* u^0 *approach* u^0 *exponentially.*

These assertions are independent of the $\sigma \in (0,1]$ *appearing in* (3.7).

The equivalence of (i) and (ii) is contained in Corollary 3.5. The equivalence of (ii) and (iii) follows from the next lemma.

LEMMA 3.7. *Let* $w \mapsto \phi(w)$ *be a differentiable map of a neighborhood* $\mathcal{W}$ *of* 0 *in a Banach space* $\mathcal{X}$ *with* $\phi(0) = 0$. *Then* $S = \phi_w(0)$ *has its spectrum strictly inside the unit disk if and only if there exist a neighborhood* $\tilde{\mathcal{W}}$ *of* 0, *a positive number* c, *and a number* $\delta \in [0,1)$ *such that the iterated maps* ϕ^k *satisfy*

$$|\phi^k(v)| \le c\delta^k \tag{3.8}$$

for $\mathbf{x} \in \tilde{\mathcal{W}}$.

Proof. The assertion about the spectrum is equivalent to the existence of j and $\delta < 1$ such that $\|S^j w\| \le \delta \|w\|$. Similarly, (3.8) is equivalent to the existence of an invariant neighborhood $\tilde{W}$ of zero, of j, and of δ such that

$$\|\phi^j(w)\| \le \delta \|w\|$$

for $w \in \tilde{W}$. The assertion follows now by simple calculus considerations. There are no changes in the context of Banach manifolds. □

In the application to the proof of Theorem 3.6, ϕ is the evaluation of the semiflow $\boldsymbol{\phi}(\cdot,\cdot,\zeta)$ at the same time t.

The semigroup $S(\cdot;\zeta)$ has a generator $A(\zeta)$. We shall show that isolated eigenvalues of the generator depend differentiably on ζ. Then the conditions of the following theorem make sense.

THEOREM 3.8 (Hopf bifurcation theorem). *Let* (1.2a), (1.2b) *and Hypothesis* 1.1 *hold. Let* λ_0 *be a simple nonzero eigenvalue of* $A(0)$ *lying on the imaginary axis. Then there is an interval* (ζ_-, ζ_+) *containing* 0 *and a unique continuous function* $(\zeta_-, \zeta_+) \ni \zeta \mapsto \hat{\lambda}(\zeta)$ *such that* $\hat{\lambda}(0) = \lambda_0$ *and* $\hat{\lambda}(\zeta)$ *is an isolated simple eigenvalue of* $A(\zeta)$ *for each* $\zeta \in (\zeta_-, \zeta_+)$. *Moreover,* $\hat{\lambda} \in C^k(\zeta_-, \zeta_+)$. *Let* $v^{\pm}$ *be a basis for the eigenspace corresponding to the eigenvalues* $\pm\lambda_0$ *such that, with* $\lambda = i\mu$, $S(t)v^+ = \cos(t\mu)v^+ + \sin(t\mu)v^-$. *Suppose that:*

(i) *If there is an integer* n *such that* $n\lambda_0$ *is an eigenvalue of* $A(0)$, *then* $n = \pm 1$ (*this is the* weak nonresonance condition).

(ii) $\Re\, \lambda_\zeta(0) \neq 0$. (*This is the* transversality condition).

(iii) *There are only isolated eigenvalues of finite multiplicity of the generator on the closed right half space.* (*This is satisfied under mild assumptions on the constitutive functions.*)

Then there exist an interval $[0, \varepsilon_0)$ *and unique* C^k *maps*

$$[0, \varepsilon_0) \ni \varepsilon \mapsto \hat{u}(\cdot, \cdot; \varepsilon) \in C^{2,\alpha}_{\mathrm{P}}(\Omega \times \mathbb{R}),$$
$$\hat{T} : [0, \varepsilon_0) \to (0, \infty), \quad \hat{\zeta} : [0, \varepsilon_0) \to \mathbb{R}$$

such that $\hat{u}(\cdot, \cdot; \varepsilon)$ *and* $\hat{\zeta}(\varepsilon)$ *satisfy* (1.3), $\hat{u}(x, \cdot; \varepsilon)$ *has period* $\hat{T}(\varepsilon)$, *and*

$$\begin{gathered} \hat{T}(0) = 2\pi/|\lambda_0|, \quad \hat{T}'_\varepsilon(0) = 0, \qquad \hat{\zeta}(0) = 0, \quad \hat{\zeta}_\varepsilon(0) = 0, \\ \int_\Omega \hat{u}(\mathbf{x}, t; \varepsilon)v^+ = \varepsilon, \qquad \int_\Omega \hat{u}(\mathbf{x}, t; \varepsilon)v^- = 0. \end{gathered} \tag{3.9}$$

Moreover, every small solution of (1.3) *having period close to* $\hat{T}(0)$ *is given by a time shift of* $\hat{u}(\cdot, \cdot; \varepsilon)$.

Suppose that (i) *is replaced with the* strong nonresonance condition:

(iv) *If there is a real number* r *such that* $r\lambda_0$ *is an eigenvalue of* $A(0)$, *then* $r = \pm 1$.

Then every small periodic solution of (1.3) *is given by a time shift of* $\hat{u}(\cdot, \cdot; \varepsilon)$.

Let (ii), (iii), *and* (iv) *hold. If the trivial solution is stable for* $\zeta > 0$ *and if* $\hat{\zeta}_{\varepsilon\varepsilon}(0) > 0$ (*i.e., if the bifurcation is supercritical*), *then* $\hat{u}(\cdot, \cdot; \varepsilon)$ *is unstable. If the trivial solution is stable for* $\zeta > 0$ *and if* $\hat{\zeta}_{\varepsilon\varepsilon}(0) < 0$ (*i.e., if the bifurcation is subcritical*), *then* $\hat{u}(\cdot, \cdot; \varepsilon)$ *is stable. The remaining two cases are analogous.*

There are many normalizations for $u(\cdot, \cdot, \varepsilon)$ equivalent to (3.19), with the equivalence justified by the implicit-function theorem. The normalization (3.19) says that the orthogonal projection of $\tilde{u}(\cdot, \cdot, \varepsilon)$ onto the space spanned by v^+ and v^- lies on a given line parametrized by ε.

In order to carry out our analysis and to use Theorems 3.6 and 3.8, we need some general information about the nature of the spectrum. The spectrum in the right half space can be controlled under natural weak assumptions on the coefficients by using the regularity results of [1] together with the next proposition. To treat specific problems one requires detailed information about the point spectrum. This information is obtained for (2.1) in [7].

Consider the problem

$$(\mathbf{G} + \theta\mathbf{H}) : v_{\mathbf{xx}} = 0 \quad \text{in } \Omega, \tag{3.10a}$$
$$(\mathbf{b} + \theta\mathbf{a}) \cdot v_{\mathbf{x}} = 0 \quad \text{on } \Gamma. \tag{3.10b}$$

Let $\mathcal{R}$ be the unbounded component of the subset of $\mathbb{C}$ consisting of all θ such that problem (3.10) is elliptic in the sense of Agmon, Douglis, and Nirenberg.

PROPOSITION 3.9. *If $\theta \in \mathcal{R}$, then θ is also in $\mathcal{R}$ for small perturbations of the coefficients. There are only isolated eigenvalues of finite multiplicity of the generator of the semigroup defined by* (3.5) *in* $\mathcal{R}$.

For completeness we present an unusual feature of the semiflow, which we need only for the linear equation. It takes, however, only a slight additional effort to prove the following compactness result also for the nonlinear problem. Consider the problem

$$F(\mathbf{x}, u, u_{\mathbf{x}}, u_{\mathbf{xx}}, u_t, u_{t\mathbf{x}}, u_{t\mathbf{xx}}, \zeta) = \mu u_t \qquad \text{in } \Omega_T, \tag{3.11a}$$

$$B(\mathbf{x}, u, u_{\mathbf{x}}, u_t, u_{t\mathbf{x}}, \zeta) = 0 \qquad \text{on } \Gamma_T, \tag{3.11b}$$

$$u(\cdot, 0) = u_0(\cdot) \qquad \text{in } \Omega, \tag{3.11c}$$

where the term μu_t on the right-hand side is of minor importance. It is needed to prevent 0 from being in the spectrum of an operator with compact resolvent.

For $u_0 \in C^{2,\sigma}$ and fixed $\zeta \in \mathcal{I}$ there exists μ such that (3.11) defines a local flow $\boldsymbol{\psi}$ on $C^{2,\sigma}$ and $h^{2,\sigma}$ in a neighborhood of u_0.

THEOREM 3.10. *The map*

$$\mathcal{H}^{2,\sigma}(\zeta) \ni (u_0, u_1) \mapsto (\boldsymbol{\phi}(t, u_0, u_1, \zeta) - (\boldsymbol{\psi}(t, u_0, \zeta), \boldsymbol{\psi}_t(t, u_0, \zeta))) \in (C^{2,\sigma})^2 \tag{3.12}$$

is compact for $t > 0$ whenever it is defined. More precisely, this map is the composition of a continuous and differentiable nonlinear map and a compact linear inclusion, which maps bounded sets to precompact sets.

Remark. Observe that $\dot{\boldsymbol{\phi}}$ is defined for negative time. Theorem 3.10 implies that $\boldsymbol{\phi}$ is a compact perturbation of a noncompact map. This implies that the semiflow is not smoothing.

Now suppose for simplicity and essentially without loss of generality that $u = 0$ satisfies (3.11) with zero initial data. The linearization of (3.11) about $u = 0$ is

$$\mathbf{G} : v_{\mathbf{xx}} + \mathbf{g} \cdot v_{\mathbf{x}} + \gamma v + \mathbf{H} : v_{\mathbf{xx}t} + \mathbf{h} \cdot v_{\mathbf{x}t} + \eta v_t = \mu v_t \quad \text{in } \Omega_T, \tag{3.13a}$$

$$\mathbf{a} \cdot v_{\mathbf{x}} + \alpha v + \mathbf{b} \cdot v_{\mathbf{x}t} + \beta v_t = 0 \qquad \text{on } \Gamma_T, \tag{3.13b}$$

$$v(\cdot, 0) = 0 \qquad \text{in } \Omega. \tag{3.13c}$$

The coefficients depend on $\mathbf{x}$ but not on t. We shall see that (3.13) defines an analytic group on $h^{2,\sigma}$ (and on $C^{2,\sigma}$). Hence the spectrum of the group is determined by the spectrum of the generator. On the other hand, this group describes the parabolic-hyperbolic semigroup up to a compact perturbation. Hence we can control the essential spectrum of the parabolic-hyperbolic semigroup provided we can control the spectrum of the generator of the group defined by (3.13), i.e., if we can control $\mathcal{R}$. This can be done under conditions that can often be readily verified.

THEOREM 3.11. *Let Hypothesis* 1.1 *hold. Suppose that* (i) *F has divergence form, i.e., that there is a vector-valued function $\mathbf{f}$ and a function M such that*

$$F = \partial_{\mathbf{x}} \cdot \mathbf{f}(u, u_{\mathbf{x}}, u_t, u_{t\mathbf{x}}) + M(u, u_{\mathbf{x}}, u_t, u_{t\mathbf{x}}, \zeta),$$

(ii) *$\mathbf{G}$ is uniformly positive-definite for all $\mathbf{x}$, and* (iii) *the boundary function B has the form*

$$B(\mathbf{x}, u, u_{\mathbf{x}}, u_t, u_{t\mathbf{x}}, \zeta) = \boldsymbol{\nu} \cdot \mathbf{f}(\mathbf{x}, u, u_{\mathbf{x}}, u_t, u_{t\mathbf{x}}, \zeta) + H(\mathbf{x}, u, u_t, u_{\mathbf{x}}, \zeta), \tag{3.14}$$

where H depends only on tangential components of $u_{\mathbf{x}}$. Then $\mathcal{R}$ contains the right closed half plane $\{\mu \in \mathbb{C} : \Re\, \mu \geq 0\}$. In particular, hypothesis (iii) *in Theorem* 3.8 *is satisfied.*

As mentioned above, there are important differences between our work and that of Da Prato and Lunardi [10]: We do not prove a center-manifold theorem, the space of admissible initial data is not a linear space, and the equations are different. The obstacle to using the techniques of [10], Potier-Ferry [21], [22], Renardy [23], and Xu and Marsden [25] is that the space of admissible initial data is not a linear space. We shall return to this point in the discussion of section 8.

We complete this introduction by giving an outline of the paper. We start with a study of linear equations in section 4 and continue with the manifold structure of $\mathcal{H}^{2,\sigma}$ in section 5. Section 6 is devoted to well-posedness questions. In section 7 we state, prove, and apply an abstract Hopf bifurcation theorem. In section 8, we check that our physical examples have the requisite properties.

4. The linear equation. In this section we obtain bounds and regularity results for the linearized parabolic-hyperbolic equation by exploiting estimates for associated linear parabolic equations. Then we study the spectrum of the generator of the semigroup for our linearized problem, which leads to the treatment of associated linear elliptic problems and of a local flow closely related to the local semiflow we intend to study.

The linear parabolic problem. We consider the linear parabolic problem

$$u_t - \mathbf{H} : u_{\mathbf{xx}} - \mathbf{h} \cdot u_{\mathbf{x}} - \eta u = f \qquad \text{in } \Omega_T, \tag{4.1a}$$

$$\mathbf{b} \cdot u_{\mathbf{x}} + \beta u = g \qquad \text{in } \Gamma_T, \tag{4.1b}$$

$$u(\cdot, 0) = u_0(\cdot) \quad \text{on } \Omega \tag{4.1c}$$

with $\mathbf{H}, \mathbf{h}, \eta \in C^{\sigma}_{\mathrm{P}}$, $\mathbf{b}, \beta \in C^{1,\sigma}$, $\mathbf{H}$ uniformly positive-definite, and $\mathbf{b}$ transversal at the boundary. We first state a standard existence result.

THEOREM 4.1. *Let $u_0 \in C^{2,\sigma}(\Omega)$, $g \in C^{1,\sigma}_{\mathrm{P}}(\Gamma_T)$, and $\mathbf{b} \cdot u_{0\mathbf{x}} + \beta u_0 = g(\cdot, 0)$ for $\mathbf{x} \in \Gamma$. Then there exist a constant $c > 0$ and a unique classical solution $u \in C^{2,\sigma}_{\mathrm{P}}(\Omega_T)$ of* (4.1) *that satisfies*

$$\|u\|_{C^{2,\sigma}_{\mathrm{P}}(\Omega_T)} \le c \left(\|u_0\|_{C^{2,\sigma}(\Omega)} + \|f\|_{C^{\sigma}_{\mathrm{P}}(\Omega_T)} + \|g\|_{C^{1,\sigma}_{\mathrm{P}}(\Gamma_T)} \right). \tag{4.2}$$

The constant remains bounded if T tends to zero.

This type of estimate is well known and can essentially be found in Friedman [11], Ladyženskaja, Ural'ceva, and Solonnikov [18], and Solonnikov [24]. The approach of the first two books could be applied to our problem, although the results stated there do not directly cover Theorem 4.1.

To justify (4.2) without technical difficulty, we make use of the fact that (4.1) is a scalar equation. (Nevertheless, similar results hold for Petrovskii parabolic systems [24], and all our results are readily extended to these systems. The algebraic conditions for these systems, especially the complementing condition, are natural but quite complicated in general.)

Sketch of the proof. First we consider the whole-space problem with frozen coefficients. After a linear transformation and a neglect of the lower-order terms, (4.1a) reduces to the pure heat equation, which is obviously parabolic. For the complementing condition we flatten the boundary in the neighborhood of an arbitrary point, freeze the coefficients, and apply affine transformations. We thus obtain the problem

$$u_t - \Delta u = 0 \quad \text{in } \{\mathbf{x} : x_n < 0\}, \tag{4.3a}$$

$$u_n - d u_1 = 0 \quad \text{in } \{\mathbf{x} : x_n = 0\}, \tag{4.3b}$$

where $d \in \mathbb{R}$.

Let $u = q(x_n)\exp(i\sum_{j=1}^{n-1} x^j\xi_j + \tau t)$ be a nonconstant solution to (4.3). Then $u = \text{const}\,\exp(i\sum_{j=1}^{n-1} x^j\xi_j + \rho x_n + \tau t)$ with $\tau = -|\xi|^2 + \rho^2$ and $\rho = id\xi_1$. Hence $\tau = -|\xi|^2 - d^2\xi_1^2 < 0$ and the complementing condition of [24] is satisfied for problem (4.1). Theorem 4.1 is now a very special case of Theorem 4.9 of [24]. □

Having estimates and existence results for C^σ or for h^σ is more or less equivalent. More precisely, suppose that L is a linear operator from C^σ to itself for $0 < \sigma < 1$. Then L maps h^σ to itself because h^σ is the closure of $C^{\sigma+\varepsilon}$ in C^σ, since L maps $C^{\sigma+\varepsilon}$ to itself (and hence to h^σ), and finally because a C^σ estimate implies immediately an a priori estimate in h^σ.

Let us now assume that L maps h^σ to itself for $0 < \sigma < 1$. Clearly $C^\sigma \subset h^{\sigma-\varepsilon}$. Let $f \in C^\sigma$. We approximate f in $h^{\sigma-\varepsilon}$ by f_j such that $\|f_j\|_{C^\sigma} \le c\|f\|_{C^\sigma}$. Then $\|Lf_j\|_{C^\sigma} = \|Lf_j\|_{h^\sigma} \le c\|f_j\|_{h^\sigma} = c\|f_j\|_{C^\sigma}$, which is uniformly bounded. Moreover, Lf_j converges to Lf in $C^{\sigma-\varepsilon}$. Hence $\|Lf\|_{C^\sigma} \le c\|f\|_{C^\sigma}$. Below we use this simple observation without mention.

The semigroup defined by the parabolic-hyperbolic equation. We turn to the linear parabolic-hyperbolic problem

$$u_{tt} - \mathbf{G} : u_{\mathbf{xx}} - \mathbf{g}\cdot u_{\mathbf{x}} - \gamma u - \mathbf{H} : u_{\mathbf{xx}t} - \mathbf{h}\cdot u_{\mathbf{x}t} - \eta u_t = w \qquad \text{in } \Omega_T, \tag{4.4a}$$

$$\mathbf{a}\cdot u_{\mathbf{x}} + \alpha u + \mathbf{b}\cdot u_{\mathbf{x}t} + \beta u_t = g \qquad \text{in } \Gamma_T, \tag{4.4b}$$

$$u(\cdot,0) = u_0(\cdot) \quad \text{in } \Omega, \tag{4.4c}$$

$$u_t(\cdot,0) = u_1(\cdot) \quad \text{in } \Omega \tag{4.4d}$$

with coefficients depending on $\mathbf{x}$ but not on t. Considering (4.4a) as a parabolic equation for u_t, we infer from Theorem 4.1 the following a priori estimate for $u \in Y^{2,\sigma}$:

$$\|u_t\|_{C_{\mathrm{P}}^{2,\sigma}(\Omega_T)} \le c\left(\|u_1\|_{C^{2,\sigma}(\Omega)} + \|w\|_{C_{\mathrm{P}}^{\sigma}} + \|g\|_{C_{\mathrm{P}}^{1,\sigma}} + \|u\|_{C_{\mathrm{P}}^{2,\sigma}}\right). \tag{4.5}$$

Since

$$\begin{aligned}\|u(t)\|_{C^{2,\sigma}(\Omega)} &\le \|u_0\|_{C^{2,\sigma}} + \left\|\int_0^t u_t(\cdot,\tau)d\tau\right\|_{C_{\mathrm{P}}^{2,\sigma}(\Omega_t)}\\ &\le \|u_0\|_{C^{2,\sigma}} + c\int_0^t \|u_t(\tau)\|_{C^{2,\sigma}(\Omega)}d\tau\end{aligned} \tag{4.6}$$

and since $u(\cdot,t+s) - u(\cdot,t) = \int_t^{t+s} u_t(\cdot,\tau)\,d\tau$, we obtain

$$\begin{aligned}\|u_t\|_{C_{\mathrm{P}}^{2,\sigma}} \le c\Big(&\|u_0\|_{C_{\mathrm{P}}^{2,\sigma}} + \|u_1\|_{C_{\mathrm{P}}^{2,\sigma}} + \|w\|_{C_{\mathrm{P}}^{\sigma}} + \|g\|_{C_{\mathrm{P}}^{1,\sigma}}\\ &+ \int_0^T \|u_t\|_{C^{2,\sigma}(\Omega_\tau)}d\tau + T^{1-\sigma/2}\|u_t\|_{C_{\mathrm{P}}^{2,\sigma}(\Omega_T)}\Big).\end{aligned} \tag{4.7}$$

For T sufficiently small, Gronwall's inequality now implies that

$$\|u_t\|_{C_{\mathrm{P}}^{2,\sigma}(\Omega_T)} \le c(\|u_0\|_{C^{2,\sigma}(\Omega)} + \|u_1\|_{C^{2,\sigma}(\Omega)} + \|w\|_{C_{\mathrm{P}}^{\sigma}(\Omega_T)} + \|g\|_{C_{\mathrm{P}}^{1,\sigma}(\Gamma_T)})\,. \tag{4.8}$$

This motivates the definition of the space $Y^{2,\sigma}$. The a priori estimate (4.8) can be used to get the following well-posedness result for the linear equation, at first for small t and then, by iteration, for arbitrary t. We omit the proof because the arguments

are standard and because we shall obtain this result as a special case of a result for a nonlinear problem.

PROPOSITION 4.2. *Suppose that* $u_0, u_1 \in C^{2,\sigma}$, $w \in C^\sigma$, *and* $g \in C^{1,\sigma}$, *and that*

$$\mathbf{a} \cdot u_{0\mathbf{x}} + \alpha u_0 + \mathbf{b} \cdot u_{1\mathbf{x}} + \beta u_1 = g(\cdot, 0) \quad \text{on } \Gamma. \tag{4.9}$$

Then there exists a unique solution $u \in Y^{2,\sigma}$ *of* (4.5), *which satisfies*

$$\|u\|_{Y^{2,\sigma}} \le ce^{ct}(\|u_0\|_{C^{2,\sigma}} + \|u_1\|_{C^{2,\sigma}} + \|w\|_{C^\sigma_{\mathrm{P}}} + \|g\|_{C^{1,\sigma}_{\mathrm{P}}}). \tag{4.10}$$

Proposition 4.2 can be applied in the special case that $w \equiv 0$ and $g \equiv 0$. The map $(u_0, u_1, t) \mapsto (u(t), u_t(t))$ is a semigroup $S(\cdot)$ on the subspace $\mathcal{H}_1^{2,\sigma}(\zeta)$ of $C^{2,\sigma}(\Omega) \times C^{2,\sigma}(\Omega)$ consisting of those functions satisfying (4.9) with $g = 0$. We shall see that this semigroup is analytic, i.e., there is a bounded holomorphic extension of the time variable to a sector in the complex plane.

Since regularity is not an issue at the moment, we suppose that u is sufficiently smooth and satisfies (4.4) with $w = 0$, $g = 0$, and coefficients independent of t. Then $u_{tt} \in C^\sigma_{\mathrm{P}}$. The function $v = tu_{tt}$ satisfies

$$v_t - \mathbf{H} : v_{\mathbf{xx}} - \mathbf{h} \cdot v_{\mathbf{x}} - \eta v = u_{tt} + t(\mathbf{G} : u_{\mathbf{xx}t} + \mathbf{g} \cdot u_{\mathbf{x}t} + \gamma u_t) \tag{4.11}$$

with similar boundary conditions. It follows that there is a function c depending continuously on t such that

$$\|v\|_{C^{2,\sigma}_{\mathrm{P}}} \le c\|u\|_{Y^{2,\sigma}} \le c\left(\|u_0\|_{C^{2,\sigma}} + \|u_1\|_{C^{2,\sigma}}\right), \tag{4.12}$$

$$\|u_{tt}(t)\|_{C^{2,\sigma}} \le ct^{-1}\left(\|u_0\|_{C^{2,\sigma}} + \|u_1\|_{C^{2,\sigma}}\right). \tag{4.13}$$

These a priori estimates can easily be turned into an estimate for all solutions. Proposition 4.2 implies that problem (4.4) with coefficients independent of t defines a semigroup on $\mathcal{H}_1^{2,\sigma}$, which is analytic by (4.13) and the obvious estimate $\|u_t(t)\|_{C^{2,\sigma}} \le c$.

As mentioned above, Proposition 4.2 depends on Petrovskii parabolicity with respect to $\dot{u}$, which is not hard to check also for Example 2.3.

The spectrum. We turn to the spectrum of the semigroup defined by problem (4.4). In the first step, for given u, we consider the following elliptic problem for v:

$$\mathbf{H} : v_{\mathbf{xx}} + \mathbf{h} \cdot v_{\mathbf{x}} + \eta v - \mu v = -\mathbf{G} : u_{\mathbf{xx}} - \mathbf{g} \cdot u_{\mathbf{x}} - \gamma u - w \quad \text{in } \Omega_T, \tag{4.14a}$$

$$\mathbf{a} \cdot v_{\mathbf{x}} + \alpha v = -\mathbf{b} \cdot u_{\mathbf{x}} - \beta u + g \quad \text{in } \Gamma_T. \tag{4.14b}$$

PROPOSITION 4.3. *Suppose that the real part of* μ *is sufficiently large. Then*

$$\|v\|_{C^{2,\sigma}} \le c\left(\|u\|_{C^{2,\sigma}} + \|w\|_{C^\sigma} + \|g\|_{C^{1,\sigma}}\right), \tag{4.15}$$

and $v \in C^{2,\sigma}$ *if* $u \in C^{2,\sigma}$ *and* $w \in C^\sigma$.

Proof. It is clear that $v \in C^{2,\sigma}$ if $u \in C^{2,\sigma}$. The parabolic equation (4.1) generates a semigroup. Hence μ is in the resolvent set of the operator

$$u \mapsto \mathbf{H} : u_{\mathbf{xx}} + \mathbf{h} \cdot u_{\mathbf{x}} + \eta u \tag{4.16}$$

on $\{u : \mathbf{a} \cdot u_{\mathbf{x}} + \alpha u = 0\}$ if its real part is sufficiently large. Hence we obtain (4.15) by Schauder estimates. It is again not hard to see that this result holds for systems and in particularly it applies to systems with coefficients obtained from linearizing Example 2.3. □

We use Proposition 4.3 to reduce the problem (3.13) for a partial differential equation into one for an ordinary differential equation on $h^{2,\sigma}$. By the implicit-function theorem we can solve the next two equations.

$$F(\mathbf{x}, u, u_{\mathbf{x}}, u_{\mathbf{xx}}, v, v_{\mathbf{x}}, v_{\mathbf{xx}}, \zeta) - \mu v = 0 \quad \text{in } \Omega, \tag{4.17a}$$

$$B(\mathbf{x}, u, u_{\mathbf{x}}, v, v_{\mathbf{x}}, \zeta) = 0 \quad \text{on } \Gamma, \tag{4.17b}$$

$$u_t = v \tag{4.17c}$$

for v in terms of u and ζ if u is in the neighborhood of a given function $\tilde{u}$ since the derivative at $\tilde{u}$ is an invertible map if μ is sufficiently large. Hence we may rewrite problem (4.17) in the form

$$u_t = \Psi(u) \tag{4.18}$$

with Ψ a k-times differentiable vector field in the neighborhood of the given function u_0. The following proposition is now obvious.

PROPOSITION 4.4. *Problem* (4.17) *defines a k-times differentiable local flow in a neighborhood $\mathcal{U} \subset C^{2,\sigma}$ of u_0.*

Similar but simpler arguments apply to the linear problem

$$\mathbf{G} : u_{\mathbf{xx}} + \mathbf{g} \cdot u_{\mathbf{x}} + \gamma u + \mathbf{H} : u_{t\mathbf{xx}} + \mathbf{h} \cdot u_{t\mathbf{x}} + \eta u_t = \mu u_t - w \quad \text{in } \Omega_T, \tag{4.19a}$$

$$\mathbf{a} \cdot u_{\mathbf{x}} + \alpha u + \mathbf{b} \cdot u_{t\mathbf{x}} + \beta u_t = f \quad \text{in } \Gamma_T, \tag{4.19b}$$

$$u(\cdot, 0) = u_0(\cdot) \quad \text{on } \Omega. \tag{4.19c}$$

This problem defines a group $\hat{S}$ on $C^{2,\sigma}$ and on $h^{2,\sigma}$ if $w = 0$ and $f = 0$, implicitly parametrized by ζ. The solution of (4.19) satisfies

$$\|u(t)\|_{C^{2,\sigma}} \leq c \left(\|u_0\|_{C^{2,\sigma}} + \int_0^t (\|w(\tau)\|_{C^\sigma(\Omega))} + \|f(\tau)\|_{C^{1,\sigma}(\Gamma)}) \, d\tau \right). \tag{4.20}$$

Let u be the solution to problem (4.4). Then $v(\cdot, t) = u(\cdot, t) - \hat{S}(t)u_0$ satisfies

$$\mathbf{G} : v_{\mathbf{xx}} + \mathbf{g} \cdot v_{\mathbf{x}} + \gamma v + \mathbf{H} : v_{t\mathbf{xx}} + \mathbf{h} \cdot v_{t\mathbf{x}} + \eta v_t = \mu u_t + u_{tt} \quad \text{in } \Omega_T, \tag{4.21a}$$

$$\mathbf{a} \cdot v_{\mathbf{x}} + \alpha v + \mathbf{b} \cdot v_{t\mathbf{x}} + \beta v_t = 0 \quad \text{in } \Gamma_T, \tag{4.21b}$$

$$v(\cdot, 0) = 0 \quad \text{on } \Omega. \tag{4.21c}$$

From (4.20) we obtain

$$\|v(t)\|_{C^{2,\sigma}} + \|v_t(t)\|_{C^{2,\sigma}} \leq c \int_0^t (\|u_{tt}\|_{C^\sigma} + \|u_t\|_{C^\sigma}) \, d\tau. \tag{4.22}$$

We have seen that the semigroup S defined by Proposition 4.2 is analytic; hence

$$u_{tt}(t) \in C^{2,\sigma}, \tag{4.23}$$

$$\|u_{tt}(t)\|_{C^{2,\sigma}} \leq ct^{-1} \left(\|u_0\|_{C^{2,\sigma}} + \|u_1\|_{C^{2,\sigma}} \right). \tag{4.24}$$

Thus, the map

$$\mathcal{H}_1^{2,\sigma} \ni (u_0, u_1) \mapsto u_{tt} \in L^1([0, T], C^\sigma(\Omega)) \tag{4.25}$$

is compact. Similarly, the map to u_t is compact.

For later use we formulate our conclusion, which we obtain now from (4.20):

PROPOSITION 4.5. *The map*

$$\mathcal{H}^{2,\sigma} \ni (u_0, u_1) \mapsto ((\hat{S}(t)u_0, \hat{S}_t(t)u_0) - S(t)[u_0, u_1]) \in (C^{2,\sigma})^2 \tag{4.26}$$

is compact.

The same arguments provide a proof of Theorem 3.10. Proposition 4.5 yields the first half of the control of the spectrum: If we can control the spectrum of the generator of $\hat{S}$ defined by (4.19), then we can also control the spectrum of the generator of $S(t)$. That control is provided by the Theorem 3.11.

Proof of Theorem 3.11. The notion of ellipticity of [1] does not depend on lower-order terms. The assertion of Theorem 3.11 is equivalent to the following: Suppose that $\Re\,\lambda \geq 0$ and that $\mu \in \mathbb{R}$ is sufficiently large. Then

$$(\mathbf{G} + \lambda\mathbf{H}) : u_{\mathbf{xx}} - \mu u = 0 \quad \text{in } \Omega, \tag{4.27a}$$

$$(\mathbf{G} + \lambda\mathbf{H}) : \boldsymbol{\nu} u_{\mathbf{x}} + \mathbf{d} \cdot u_{\mathbf{x}} = 0 \quad \text{on } \Gamma \tag{4.27b}$$

is elliptic, where $\mathbf{d} = \partial h/\partial u_{\mathbf{x}}$ is a tangential vector field.

We obtain the obvious energy estimate

$$\mu \int_\Omega u\bar{u} + \int_\Omega (\mathbf{G} + \lambda\mathbf{H}) : u_{\mathbf{x}}\bar{u}_{\mathbf{x}} + \int_\Gamma \mathbf{d} \cdot u_{\mathbf{x}}\bar{u} = 0. \tag{4.28}$$

Taking the real part of (4.28), integrating it by parts on the boundary, and using trace estimates, we get

$$\mu \int_\Omega u\bar{u} + \int_\Omega \mathbf{G} : u_{\mathbf{x}}\bar{u}_{\mathbf{x}} \leq c\|u\|_2\|u_{\mathbf{x}}\|_2. \tag{4.29}$$

This implies the L^2-estimates provided that μ is sufficiently large.

In particular, we can exclude the possibility of growing exponential solutions for the half space problem with constant coefficients, and hence the ellipticity and complementing conditions of [1] are satisfied. Thus problem (4.27) is elliptic for λ in the closed right half plane. □

The assumptions are satisfied in Example 2.1 provided that $\sigma_a(a, b, x) > 0$ everywhere, and are satisfied in Example 2.2 provided that (the symmetric part of) $\begin{pmatrix} \partial\sigma_1/\partial p_1 & \partial\sigma_1/\partial p_2 \\ \partial\sigma_2/\partial p_1 & \partial\sigma_2/\partial p_2 \end{pmatrix}$ is positive-definite. The arguments apply also to Example 2.3 provided that (the symmetric part of)

$$\begin{pmatrix} N_\nu & N_\eta & N_\mu \\ H_\nu & H_\eta & H_\mu \\ M_\nu & M_\eta & M_\mu \end{pmatrix}$$

is positive-definite. Incidentally, it is interesting to note that global existence theorems do not require this definiteness.

5. The manifold of admissible initial data.

Proof of Theorem 3.2. We suppress the dependence on ζ. By standard trace theory there is a linear map

$$J : C^{2,\sigma}(\Gamma) \times C^{1,\sigma}(\Gamma) \to C^{2,\sigma}(\Omega) \tag{5.1}$$

which has the properties

$$J(u_0,u_1)|_\Gamma = u_0, \tag{5.2}$$

$$\boldsymbol{\nu}\cdot J(u_0,u_1)_{\mathbf{x}|_\Gamma} = u_1. \tag{5.3}$$

Assumption 1.1 allows us to take

$$B(\mathbf{x},u_0,u_{0\mathbf{x}},u_1,u_{1\mathbf{x}}) = \boldsymbol{\nu}\cdot u_{0\mathbf{x}t} - \tilde{B}(\mathbf{x},u_0,u_{0\mathbf{x}},u_1,u_{1\mathbf{x}}), \tag{5.4}$$

where only tangential components of $u_{1\mathbf{x}}$ enter into $\tilde{B}$.

Let $\mathcal{X} = (C^{2,\sigma}(\Omega))^2$, $\mathcal{Y} = C^{1,\sigma}(\Gamma)$, $\psi(u_0,u_1) := B(\mathbf{x},u_0,u_{0,\mathbf{x}},u_1,u_{1,\mathbf{x}})$, and $L(z) := (0,J(0,z))$. We easily verify that $\psi \in C^k(\mathcal{X},\mathcal{Y})$, L has closed range, and $z = \psi'(u_0,u_1)[L(z)]$. This implies that the set $E = \psi^{-1}(\{0\})$ is a C^k-submanifold of $C^{2,\sigma}$. It is not hard to see that this construction has the requisite dependence on ζ. Moreover, the argument carries over without the special choice of (5.4), because any oblique boundary condition can be written in this form. Finally, we obtain the same assertions in the spaces $h^{2,\sigma}$. □

The key property of the problem used in this proof is the representation (5.4). The obvious generalization to systems is clearly satisfied in Example 2.3.

6. Well-posedness.

Proofs of Theorems 3.1 *and* 3.3. We want to find solutions to (1.3) for $(u_0,u_1) \in \mathcal{H}^{2,\sigma}$. Define the operators $\mathfrak{A}$ and $\mathfrak{B}$ by

$$\mathfrak{A}v = F_{\dot{\mathbf{P}}} : v_{\mathbf{xx}}, \quad \mathfrak{B}v = B_{\dot{\mathbf{p}}}\cdot v_{\mathbf{x}}, \tag{6.1a,b}$$

where the derivatives of F and B are evaluated at (u_0,u_1). We rewrite (1.3) as

$$w_t - \mathfrak{A}w = G(u,w) \quad \text{in } \Omega_T, \tag{6.2a}$$

$$\mathfrak{B}w = H(u,w) \quad \text{on } \Gamma_T, \tag{6.2b}$$

$$w(\cdot,0) = 0 \quad \text{in } \Omega, \tag{6.2c}$$

where $u(x,t) = u_0(x) + \int_0^t [w(x,\tau)+u_1(x)]d\tau$, and

$$G(u,w) = F(\mathbf{x},u,u_{\mathbf{x}},u_{\mathbf{xx}},u_1+w,(u_1+w)_{\mathbf{x}},(u_1+w)_{\mathbf{xx}}) - \mathfrak{A}w, \tag{6.3a}$$

$$H(u,w) = B(\mathbf{x},u,u_{\mathbf{x}},u_1+w,(u_1+w)_{\mathbf{x}}) - \mathfrak{B}w. \tag{6.3b}$$

The main property of G and B is that they are quadratic in the highest-order derivatives.

Obviously, solving (1.3) is equivalent to solving (6.2). The following estimates require no more than an inspection of the difference quotients involved.

LEMMA 6.1. *Let* $u = u_0 + tu_1 + \int_0^t w\,d\tau$ *and* $\bar{u} = u_0 + tu_1 + \int_0^t \bar{w}\,d\tau$. *Then the following estimates hold with* c *depending on* $\|u_0\|_{C^{2,\sigma}}$, $\|u_1\|_{C^{1,\alpha}}$, $\|w_{\mathbf{xx}}\|_{L^\infty}$ *and* $\|\bar{w}_{\mathbf{xx}}\|_{L^\infty}$:

$$\begin{aligned}\|G(u,w)-G(u,\bar{w})\|_{C^\sigma_{\mathrm{P}}} &+ \|H(u,w)-H(u,\bar{w})\|_{C^{1,\sigma}_{\mathrm{P}}}\\ &\le c\|(w-\bar{w})_{\mathbf{xx}}\|_\infty\left(\|w\|_{C^{2,\sigma}_{\mathrm{P}}}+\|\bar{w}\|_{C^{2,\sigma}_{\mathrm{P}}}\right)\\ &\quad + c\left(\|w_{\mathbf{xx}}\|_\infty+\|\bar{w}_{\mathbf{xx}}\|_\infty\right)\|w-\bar{w}\|_{C^{2,\sigma}_{\mathrm{P}}},\end{aligned} \tag{6.4a}$$

$$\begin{aligned}\|G(u,w)-G(\bar{u},w)\|_{C^\sigma_{\mathrm{P}}} &+ \|\ H(u,w)-H(\bar{u},w)\|\|_{C^{1,\sigma}_{\mathrm{P}}}\\ &\le c\Big(\|(u-\bar{u})_{\mathbf{xx}}\|_\infty\|w\|_{C^{2,\sigma}_{\mathrm{P}}(\Omega_T)}+\|u-\bar{u}\|_{C^{2,\sigma}_{\mathrm{P}}(\Omega_T)}\Big).\end{aligned} \tag{6.4b}$$

We construct the solution (more precisely, its time derivative) as a fixed point of the map J from $\tilde{w}$ to the solution w of

$$w_t - \mathfrak{A}w = G(\tilde{u}, \tilde{w}) \quad \text{in } \Omega_T, \tag{6.5a}$$

$$\mathfrak{B}w = H(\tilde{u}, \tilde{w}) \quad \text{on } \Gamma_T, \tag{6.5b}$$

$$w(\cdot, 0) = 0 \quad \text{in } \Omega, \tag{6.5c}$$

where $\tilde{u} = u_0 + \int_0^t (\tilde{w} + u_1)\, d\tau$. By Theorem 4.1 and Lemma 6.1 (with $\bar{w} = 0$) we have

$$\|w\|_{C^{2,\sigma}_{\mathrm{P}}} \le c(t^{\sigma/2}\|\tilde{w}\|_{C^{2,\sigma}_{\mathrm{P}}} + 1) \tag{6.6}$$

where c is uniformly bounded for $t \le 1$ if $\|\tilde{w}_{\mathbf{xx}}\|_\infty \le c\|u_1\|_{C^2}$. We choose $R = 2c$, $t = c^{-2/\sigma}$, and $Y_0 = \{w \in C^{2,\sigma}(\Omega_t) : w(\cdot, 0) = 0, \|w\|_{Y^\sigma} \le R\}$. Then J maps Y_0 to itself. Moreover, by Lemma 6.1 and Theorem 4.1,

$$\|J(w) - J(\bar{w})\|_{C^{2,\sigma}_{\mathrm{P}}} \le ct^{\sigma/2}\|w - \bar{w}\|_{C^{2,\sigma}_{\mathrm{P}}}. \tag{6.7}$$

Hence J is a contraction if t is sufficiently small. The contraction-mapping principle implies the existence of a solution.

Clearly J is k-times continuously differentiable with respect to all the data. It is differentiable at the fixed point, and the derivative with respect to w is invertible because J is a strict contraction. The implicit-function theorem implies that the fixed point w is k-times continuously differentiable with respect to initial data, constitutive functions, and parameters.

Since these statements remain true for the space $h^{2,\sigma}$, the regularity assertions about the semiflow in Theorems 3.1 and 3.3 hold. The semicontinuity of the life span $\hat{T}$ follows from the fact that if u is a solution of (1.3) in the time interval $[0, T)$ and if $0 < t < T$, then we can find a neighborhood $\mathcal{N}$ of the initial data $u(\cdot, 0), v(\cdot, 0)$ such that we can solve (1.3) up to time t for initial data in $\mathcal{N}$. This in turn is a consequence of our construction and of continuous dependence on the data. □

This proof clearly applies also to Example 2.3.

7. Hopf bifurcation. In this section we prove the Hopf bifurcation theorem by following the scheme in the original paper by Hopf [14]. Solutions of period T are solutions to

$$\mathbf{z} - \boldsymbol{\phi}(T, \mathbf{z}, \zeta) = \mathbf{0}. \tag{7.1}$$

Let $\boldsymbol{\psi}$ be a local coordinate map at $\mathbf{0}$ parametrized by ζ. Then (7.1) is equivalent to

$$\hat{\mathbf{z}} - \boldsymbol{\psi}^{-1}(\boldsymbol{\phi}(T, \boldsymbol{\psi}(\hat{\mathbf{z}}, \zeta), \zeta), \zeta) = \mathbf{0}, \tag{7.2}$$

where $\mathbf{z} = \boldsymbol{\psi}(\hat{\mathbf{z}}, \zeta)$ and $\boldsymbol{\psi}^{-1}$ refers to the inverse with respect to the first variable. We define $\hat{\boldsymbol{\phi}}(T, \hat{\mathbf{z}}, \zeta) = \boldsymbol{\psi}^{-1}(\boldsymbol{\phi}(T, \boldsymbol{\psi}(\mathbf{z}, \zeta), \zeta), \zeta)$ and drop the circumflexes and abbreviate (7.2) as (7.1). The assumptions are clearly invariant with respect to this introduction of local coordinates.

7A. Abstract Hopf bifurcation. We reformulate the Hopf bifurcation theorem in a more abstract setting. Let $\mathcal{U}$ be an open subset of the Banach space $\mathcal{X}$ and let $\mathcal{I} = \mathbb{R}$ be the parameter space. We suppose that $\boldsymbol{\phi}$ is a local parameter-dependent semiflow on the open set $\mathcal{U} \subset \mathcal{X} \times \mathcal{I}$ of class C^{k+2} with $k \ge 3$. We assume that $\mathbf{0} \in \mathcal{U}$

and that $\boldsymbol{\phi}(t, \mathbf{0}, \zeta) = \mathbf{0}$ for ζ small and $t < \hat{T}(0, \zeta)$. Then $\hat{T}(0, \zeta) = \infty$. We suppose that

$$\mathbf{S}(t, \zeta) := \boldsymbol{\phi}_{\mathbf{z}}(t, \mathbf{0}, \zeta) \tag{7.3}$$

is a (strongly) continuous semigroup.

HYPOTHESIS 7.1. *The number* 1 *is an isolated double eigenvalue of* $\mathbf{S}(T, 0)$ *with two-dimensional eigenspace* $\mathcal{X}_{\mathrm{c}}$. *There exists a basis* $(\mathbf{e}_1, \mathbf{e}_2)$ *for* $\mathcal{X}_{\mathrm{c}}$ *such that the restriction of* $\mathbf{S}(t, 0)$ *to* $\mathcal{X}_{\mathrm{c}}$ *has the matrix representation*

$$\begin{pmatrix} \cos \omega t & \sin \omega t \\ -\sin \omega t & \cos \omega t \end{pmatrix}, \tag{7.4}$$

where $T\omega = 2\pi$.

For ordinary differential equations and some partial differential equations there are center-manifold theorems which reduce the dimension of the problem. In favorable cases (for which there are no other eigenvalues on the unit circle) $\mathcal{X}_{\mathrm{c}}$ is the center subspace, hence the subscript c. The $\mathbf{S}(\cdot, \zeta)$-invariant complement is denoted $\mathcal{X}_{\mathrm{h}}$. In favorable cases $\mathcal{X}_{\mathrm{h}}$ is the hyperbolic subspace (the sum of the stable and the unstable subspaces), hence the subscript h.

Hypothesis 7.1 is a weak nonresonance condition. It implies that $\mathbf{S}(t, 0)$ has isolated simple eigenvalues $e^{\alpha(0)t}$ with $\alpha(0) = i\omega$ for $t \neq T$ in a neighborhood of T. Since isolated simple eigenvalues depend differentiably on parameters, we obtain eigenvalues $e^{\alpha(\zeta)t}$ for fixed t and a C^{k-1} function $\alpha(\zeta)$. It is clear that $e^{\alpha(\zeta)t}$ is an eigenvalue of $\mathbf{S}(t, \zeta)$ for small ζ and all $t \geq 0$.

We use the inner product $\langle a_1\mathbf{e}_1 + a_2\mathbf{e}_2, b_1\mathbf{e}_1 + b_2\mathbf{e}_2 \rangle = a_1b_1 + a_2b_2$ on $\mathcal{X}_{\mathrm{c}}$. The subspace $\mathcal{X}_{\mathrm{h}}$ is the range of $\mathbf{z} \mapsto \mathbf{z} - \mathbf{S}(T, 0)\mathbf{z}$. Let $\mathbf{P}_{\mathrm{c}}$ be the projection of $\mathcal{X}$ onto $\mathcal{X}_{\mathrm{c}}$ along $\mathcal{X}_{\mathrm{h}}$ and let $\mathbf{P}_{\mathrm{h}} = \mathbf{I} - \mathbf{P}_{\mathrm{c}}$.

HYPOTHESIS 7.2.

$$\Re \alpha_{\zeta}(0) \neq 0. \tag{7.5}$$

This hypothesis ensures that the eigenvalue of the generator crosses the imaginary axis transversally.

HYPOTHESIS 7.3. *The number* 1 *is the only point on the unit circle that lies in the spectrum of* $\mathbf{S}(T, 0)$.

HYPOTHESIS 7.4. *The spectrum of* $\mathbf{P}_{\mathrm{h}}\mathbf{S}(1, 0)\mathbf{P}_{\mathrm{h}}$ *is contained in the open unit ball.*

THEOREM 7.5 (Hopf bifurcation theorem). *Let Hypotheses* 7.1 *and* 7.2 *hold. Then there exist an interval* $[0, \varepsilon_0)$ *and unique* C^k *maps*

$$\begin{aligned} \big[0, \varepsilon_0) \ni \varepsilon \quad &\mapsto \tilde{\mathbf{z}}(\varepsilon) \in \mathcal{X}, \\ \big[0, \varepsilon_0) \ni \varepsilon \quad &\mapsto \tilde{T}(\varepsilon) \in (0, \infty), \\ \big[0, \varepsilon_0) \ni \varepsilon \quad &\mapsto \tilde{\zeta}(\varepsilon) \in \mathbb{R} \end{aligned}$$

such that $\tilde{\mathbf{z}}(\varepsilon)$, $\tilde{T}(\varepsilon)$, *and* $\tilde{\zeta}(\varepsilon)$ *satisfy* (7.1), *with*

$$\tilde{T}(0) = 2\pi/|\lambda_0|, \qquad \tilde{T}'(0) = 0, \qquad \tilde{\zeta}(0) = 0, \qquad \tilde{\zeta}'(0) = 0, \tag{7.6}$$

$$\mathbf{P}_{\mathrm{c}}\tilde{\mathbf{z}}(\varepsilon) = \varepsilon\mathbf{e}_1. \tag{7.7}$$

Moreover, if $(T, \mathbf{z}, \zeta)$ *is another solution of* (7.1) *with* T *close to* T_0, $\mathbf{z}$ *small, and* ζ *small, then there exist* ε *and* t *such that*

$$\mathbf{z} = \boldsymbol{\phi}(t, \tilde{\mathbf{z}}(\varepsilon), \tilde{\zeta}(\varepsilon)). \tag{7.8}$$

If, furthermore, Hypothesis 7.3 *holds, then every small periodic solution has a period close to* T_0 *and hence all small solutions of* (7.1) *are of the form* (7.8).

Finally suppose that Hypothesis 7.4 *holds. If the bifurcation is subcritical, then the periodic solution is stable for small* ε. *If the bifurcation is supercritical, then it is unstable for small* ε.

Theorem 3.8 is an immediate consequence of this theorem. To see this we recall first that the semiflow is defined on $\tilde{\mathcal{U}} = \{(\mathbf{z}, \zeta) : \mathbf{z} \in \mathcal{H}^{2,\sigma}(\zeta),\ \zeta \in \mathcal{I}\}$. By Theorem 3.2 there is a C^k diffeomorphism from a neighborhood $\mathcal{U}$ of $\mathbf{0}$ in $\mathcal{H}^{2,\sigma}(0) \times \mathbb{R}$ to $\tilde{\mathcal{U}}$. Using these coordinates we obtain a parameter-dependent semiflow on $\mathcal{U}$. The assumptions are invariant with respect to the introduction of local coordinates. This implies Theorem 3.8 provided that we can justify the different normalization in Theorem 3.8. This, however, is an immediate consequence of the regularity of the bifurcating branch and the regularity of the diffeomorphism.

7B. Proof of the abstract theorem.

Proof of Theorem 7.5. We make a linear change of variables so that $\mathcal{X}_\mathrm{c}$ and $\mathcal{X}_\mathrm{h}$ are invariant subspaces for $\mathbf{S}(t, \zeta)$ for small ζ. Here we lose one derivative. (This loss could be avoided by a bit more work.) We denote by $\mathbf{S}_\mathrm{h}(t, \zeta)$ the induced semigroup on $\mathcal{X}_\mathrm{h}$ and by $\mathbf{S}_\mathrm{c}(t, \zeta)$ the induced semigroup on $\mathcal{X}_\mathrm{c}$. The nonresonance condition is equivalent to the statement that 1 is in the resolvent set of $\mathbf{S}_\mathrm{h}(2\pi/\omega, 0)$. Hence 1 is in the resolvent set of $\mathbf{S}_\mathrm{h}(\pi/\omega, 0)$. It is also in the resolvent set of $\mathbf{S}_\mathrm{c}(\pi/\omega, 0)$ and in the resolvent set of $\mathbf{S}(\pi/\omega, 0)$.

The groups $\mathbf{S}(t, \zeta)$ define a linear group on the bilinear mappings $(\mathbf{z}_\mathrm{c}, \mathbf{v}_\mathrm{c}) \mapsto \mathbf{M} : \mathbf{z}_\mathrm{c}\mathbf{v}_\mathrm{c}$ from $\mathcal{X}_\mathrm{C}$ to $\mathcal{X}$ by

$$\mathbf{S}(t, \zeta)(\mathbf{M})(\mathbf{z}_\mathrm{c}, \mathbf{v}_\mathrm{c}) := \mathbf{S}(t, \zeta)\mathbf{M} : [\mathbf{S}_\mathrm{c}(-t, \zeta)\mathbf{z}_\mathrm{c}][\mathbf{S}_\mathrm{c}(-t, \zeta)\mathbf{v}_\mathrm{c}]. \tag{7.9}$$

Now

$$\mathbf{S}(\pi/\omega, 0)(\mathbf{M})(\mathbf{z}_\mathrm{c}, \mathbf{v}_\mathrm{c}) = \mathbf{S}(\pi/\omega, 0)\mathbf{M} : \mathbf{z}_\mathrm{c}\mathbf{v}_\mathrm{c}, \tag{7.10}$$

and by the considerations above, 1 is not in the spectrum of the operator induced by $\mathbf{S}(\pi/\omega, 0)$. Hence there exists a unique quadratic map $\mathbf{M}$ that satisfies

$$\begin{aligned} &\mathbf{M} : \mathbf{z}_\mathrm{c}\mathbf{z}_\mathrm{c} - \mathbf{S}(\pi/\omega, \zeta)\mathbf{M} : [\mathbf{S}_\mathrm{c}(-\pi/\omega, \zeta)\mathbf{z}_\mathrm{c}][\mathbf{S}_\mathrm{c}(-\pi/\omega, \zeta)\mathbf{z}_\mathrm{c}] \\ &= -\tfrac{1}{2}\boldsymbol{\phi}_{\mathbf{z}_\mathrm{c}\mathbf{z}_\mathrm{c}}(\pi/\omega, 0, \zeta) : [\mathbf{S}_\mathrm{c}(-\pi/\omega, \zeta)\mathbf{z}_\mathrm{c}][\mathbf{S}_\mathrm{c}(-\pi/\omega, \zeta)\mathbf{z}_\mathrm{c}]. \end{aligned} \tag{7.11}$$

We could define $\mathbf{M}$ for t in a neighborhood of the time π/ω as well as at time $\pi/k\omega$ for positive integers k by the same argument. We shall see later that $\mathbf{M}$ does not depend on t. Let

$$\tilde{\mathbf{z}} := \mathbf{z} - \mathbf{M} : \mathbf{z}_\mathrm{c}\mathbf{z}_\mathrm{c}. \tag{7.12}$$

The map $\mathbf{z} \mapsto \tilde{\mathbf{z}}$ is smooth and invertible near the origin. Let $\tilde{\boldsymbol{\phi}}$ be the local semiflow in the coordinates $\tilde{\mathbf{z}}$. Then

$$\tilde{\boldsymbol{\phi}}_{\tilde{\mathbf{z}}_\mathrm{c}\tilde{\mathbf{z}}_\mathrm{c}}(0, \pi/\omega, \zeta) = \mathbf{O} \tag{7.13}$$

because

(7.14)
$$\begin{aligned}\tilde{\boldsymbol{\phi}}(\pi/\omega,\tilde{\mathbf{z}},\zeta) &= \boldsymbol{\phi}(\pi/\omega,\tilde{\mathbf{z}}-\mathbf{M}:\tilde{\mathbf{z}}\tilde{\mathbf{z}},\zeta)+\mathbf{M}:[\mathbf{S}_{\mathrm{c}}(\pi/\omega,\zeta)\mathbf{z}_{\mathrm{c}}][\mathbf{S}_{\mathrm{c}}(\pi/\omega,\zeta)\mathbf{z}_{\mathrm{c}}]+O(|\mathbf{z}|^3)\\ &= \mathbf{S}(\pi/\omega,\zeta)(\tilde{\mathbf{z}}-\mathbf{M}:\tilde{\mathbf{z}}\tilde{\mathbf{z}})+\tfrac{1}{2}\boldsymbol{\phi}_{\mathbf{zz}}(\pi/\omega,\mathbf{0},\zeta):\tilde{\mathbf{z}}\tilde{\mathbf{z}}\\ &\quad +\mathbf{M}:[\mathbf{S}_{\mathrm{c}}(\pi/\omega,\zeta)\tilde{\mathbf{z}}_{\mathrm{c}}][\mathbf{S}_{\mathrm{c}}(\pi/\omega,\zeta)\tilde{\mathbf{z}}_{\mathrm{c}}]+O(|\tilde{\mathbf{z}}|^3).\end{aligned}$$

This implies the same property at $k\pi/\omega$, and the same arguments can be applied at times $\pi/k\omega$. On the other hand, the vanishing of the second derivatives is equivalent to (7.11). Thus we do get the same η for all rational multiples of π/ω, and by continuity, (7.13) holds for all t. We keep our new coordinates and drop the tilde in what follows.

We rewrite (7.1) as a system

$$\mathbf{z}_{\mathrm{h}}-\boldsymbol{\phi}_{\mathrm{h}}(t,\mathbf{z}_{\mathrm{c}}+\mathbf{z}_{\mathrm{h}},\zeta)=\mathbf{0}, \tag{7.15}$$

$$\mathbf{z}_{\mathrm{c}}-\boldsymbol{\phi}_{\mathrm{c}}(t,\mathbf{z}_{\mathrm{c}}+\mathbf{z}_{\mathrm{h}},\zeta)=\mathbf{0}. \tag{7.16}$$

Using the implicit-function theorem, we solve (7.15) for $\mathbf{z}_{\mathrm{h}}$ in terms of t, $\mathbf{z}_{\mathrm{c}}$, and ζ near $(T,\mathbf{0},0)$. We denote the function which we obtain in this way by $\tilde{\mathbf{z}}_h$. Then

$$\tilde{\mathbf{z}}_{\mathrm{h}}(t,\mathbf{0},\zeta)=\mathbf{0},\quad \frac{\partial\tilde{\mathbf{z}}_{\mathrm{h}}}{\partial\mathbf{z}_{\mathrm{c}}}(t,\mathbf{0},\zeta)=\mathbf{O},\quad \frac{\partial^2\tilde{\mathbf{z}}_{\mathrm{h}}}{\partial\mathbf{z}_{\mathrm{c}}^2}(t,\mathbf{0},\zeta)=\mathbf{O},$$

where the last statement depends on our choice of coordinates. We substitute the function just obtained into the second equation and rewrite it as

$$\mathbf{z}_{\mathrm{c}}-\mathbf{g}(t,\mathbf{z}_{\mathrm{c}},\zeta)=\mathbf{0}, \tag{7.17}$$

where $\mathbf{g}=\boldsymbol{\phi}_{\mathrm{c}}(t,\mathbf{z}_{\mathrm{c}}+\tilde{\mathbf{z}}_{\mathrm{h}}(t,\mathbf{z}_{\mathrm{c}},\zeta),\zeta)$. Then

$$\mathbf{g}_{t\mathbf{z}_{\mathrm{c}}}(T,\mathbf{0},0)=\partial_t\mathbf{S}_{\mathrm{c}}(T,\mathbf{0},0)=\begin{pmatrix}0&\omega\\-\omega&0\end{pmatrix}, \tag{7.18}$$

$$\mathbf{g}_{\mathbf{z}_{\mathrm{c}}\mathbf{z}_{\mathrm{c}}}(t,\mathbf{0},\zeta)=\mathbf{O}, \tag{7.19}$$

$$\mathbf{g}_{\zeta\mathbf{z}_{\mathrm{c}}}(T,\mathbf{0},0)=\partial_\zeta\mathbf{S}_{\mathrm{c}}(T,\mathbf{0},0)=\begin{pmatrix}\Re\,z&\Im\,z\\-\Im\,z&\Re\,z\end{pmatrix}, \tag{7.20}$$

where, with α as in Hypothesis 7.2,

$$z=T\alpha_\zeta(0). \tag{7.21}$$

Equality (7.20) can be obtained as follows: The differential of the map $A\mapsto e^{AT}$ has a two-dimensional null space at $\left(\begin{smallmatrix}0&\omega\\-\omega(0)&0\end{smallmatrix}\right)$ spanned by matrices of the form $\left(\begin{smallmatrix}a&0\\0&-a\end{smallmatrix}\right)$ and $\left(\begin{smallmatrix}0&b\\b&0\end{smallmatrix}\right)$ because

$$\det\begin{pmatrix}a-\lambda&b+\omega\\b-\omega&-a-\lambda\end{pmatrix}=\lambda^2+\omega^2-a^2-b^2. \tag{7.22}$$

Now

$$\frac{d}{da}\exp\begin{pmatrix}a&\omega T\\-\omega T&a\end{pmatrix}=\begin{pmatrix}T&0\\0&T\end{pmatrix}, \tag{7.23}$$

$$\frac{d}{db}\exp\begin{pmatrix} 0 & (b+\omega)T \\ -(b+\omega)T & a \end{pmatrix} = \begin{pmatrix} 0 & \omega \\ -\omega & 0 \end{pmatrix}. \tag{7.24}$$

Checking the eigenvalues gives (7.20).

We seek solutions of (7.17) of the form $\mathbf{z_C} = \varepsilon \mathbf{e}_1$. Then

$$\mathbf{e}_1 - \mathbf{h}(T, \varepsilon\mathbf{e}_1, \zeta) = \mathbf{0}, \quad \text{where} \quad \mathbf{h} = \mathbf{g}/\varepsilon. \tag{7.25}$$

Now $\mathbf{h} \in C^{k-1}$ because $\mathbf{g}(t, \mathbf{0}, \zeta) = \mathbf{0}$ and

$$\mathbf{h}(T, \mathbf{0}, 0) = \mathbf{e}_1, \tag{7.26}$$

$$\mathbf{h}_t(T, \mathbf{0}, 0) = \partial_t \mathbf{S}_{\mathrm{c}}(T, 0)\mathbf{e}_1 = \begin{pmatrix} 0 \\ -\omega \end{pmatrix}, \tag{7.27}$$

$$\mathbf{h}_\zeta(T, \mathbf{0}, 0) = \partial_\zeta \mathbf{S}_{\mathrm{c}}(T, 0)\mathbf{e}_1 = \begin{pmatrix} \Re\, z \\ \Im\, z \end{pmatrix}, \tag{7.28}$$

where $\Re\, z \neq 0$. Hence we can solve (7.26) for ζ and t in terms of ε. We obtain a parametrization of the points on the orbits of the periodic solutions by

$$\bar{\mathbf{z}}(s, \varepsilon) = \boldsymbol{\phi}(s\tilde{T}(\varepsilon), \varepsilon\mathbf{e}_1 + \tilde{\mathbf{z}}_{\mathrm{h}}(\tilde{T}(\varepsilon), \varepsilon\mathbf{e}_1, \tilde{\zeta}(\varepsilon))). \tag{7.29}$$

We observe that $\bar{\mathbf{z}}_{\varepsilon\varepsilon}(s, 0) = \mathbf{0}$ because $\partial^2 \tilde{\mathbf{z}}_{\mathrm{h}}(t, \mathbf{0}, \zeta(0))/\partial {\mathbf{z}_{\mathrm{c}}}^2 = \mathbf{O}$.

Our approach is based on the implicit-function theorem. Thus all small solutions with period close to T are of that form. Suppose that there is a sequence of periodic solutions with initial points $\mathbf{z}_i$, parameters ζ_i, and frequencies ω_i with $\mathbf{z}_i \to 0$ and $\zeta_i \to 0$. (The frequency of a solution is 2π divided by its least period.) Suppose that there were a subsequence such that $\omega_i \to \infty$. Then a multiple of $2\pi/\omega_i$ would be close to T and hence $\mathbf{z}_i$ would be of the form described by (7.8). This is a contradiction since the minimal period of these $\mathbf{z}_i$ is close to T. Thus the minimal periods are bounded from below. Let $\bar{\omega}$ be an accumulation point of the sequence ω_i. It is not hard to see that $e^{i\omega t}$ is necessarily an eigenvalue of $\mathbf{S}(t, 0)$. This implies the uniqueness statements.

7C. Stability.

Continuation of the proof of Theorem 7.5. We have constructed a smooth family of periodic solutions that satisfy

$$\boldsymbol{\phi}(\tilde{T}(\varepsilon), \tilde{\mathbf{z}}(\varepsilon), \tilde{\zeta}(\varepsilon)) = \tilde{\mathbf{z}}(\varepsilon). \tag{7.30}$$

The stability of these periodic solutions is determined by Floquet multipliers. In the following we shall relate the Floquet multiplier that is connected to the bifurcation to second derivatives of $\tilde{\zeta}$ with respect to ε and to the direction in which the eigenvalue crosses.

We differentiate identity (7.30) with respect to ε to get

$$\mathbf{0} = \mathbf{U}\tilde{\mathbf{z}}_\varepsilon - \tilde{\mathbf{z}}_\varepsilon + \boldsymbol{\phi}_t \tilde{T}_\varepsilon + \boldsymbol{\phi}_\zeta \tilde{\zeta}_\varepsilon, \tag{7.31}$$

where $\mathbf{U}(\varepsilon) := \boldsymbol{\phi}_{\mathbf{z}}(\tilde{T}(\varepsilon), \tilde{\mathbf{z}}(\varepsilon), \tilde{\zeta}(\varepsilon))$. We differentiate (7.31):

$$\begin{aligned} \mathbf{0} = {} & \boldsymbol{\phi}_{\mathbf{zz}} : \tilde{\mathbf{z}}_\varepsilon \tilde{\mathbf{z}}_\varepsilon + \mathbf{U}(\varepsilon)\tilde{\mathbf{z}}_{\varepsilon\varepsilon} - \tilde{\mathbf{z}}_{\varepsilon\varepsilon} + 2\tilde{T}_\varepsilon \boldsymbol{\phi}_{t\mathbf{z}} \tilde{\mathbf{z}}_\varepsilon + \boldsymbol{\phi}_{tt}(\tilde{T}_\varepsilon)^2 \\ & + \boldsymbol{\phi}_t \tilde{T}_{\varepsilon\varepsilon} + 2\boldsymbol{\phi}_{t\zeta}\tilde{\zeta}_\varepsilon \tilde{T}_\varepsilon + 2\tilde{\zeta}_\varepsilon \boldsymbol{\phi}_{\zeta\mathbf{z}} \tilde{\mathbf{z}}_\varepsilon + \boldsymbol{\phi}_{\zeta\zeta}(\tilde{\zeta}_\varepsilon)^2 + \boldsymbol{\phi}_\zeta \tilde{\zeta}_{\varepsilon\varepsilon}. \end{aligned} \tag{7.32}$$

This identity can be evaluated at $\varepsilon = 0$, where

$$\boldsymbol{\phi}_t = \boldsymbol{\phi}_{tt} = \boldsymbol{\phi}_\zeta = \boldsymbol{\phi}_{\zeta\zeta} = \boldsymbol{\phi}_{t\zeta} = \mathbf{0}$$

because

$$\boldsymbol{\phi}(t, \mathbf{0}, \zeta) = \mathbf{0}, \quad \tilde{\mathbf{z}}_\varepsilon = \mathbf{e}_1, \quad \tilde{\mathbf{z}}_{\varepsilon\varepsilon} = \mathbf{0}$$

by construction, and where

$$\boldsymbol{\phi}_{\mathbf{zz}} : \tilde{\mathbf{z}}_\varepsilon \tilde{\mathbf{z}}_\varepsilon = \boldsymbol{\phi}_{\mathbf{zz}} : \mathbf{e}_1\mathbf{e}_1 = \mathbf{0}, \qquad \tilde{T}_\varepsilon \boldsymbol{\phi}_{t\mathbf{z}} \tilde{\mathbf{z}}_\varepsilon = \omega \tilde{T}_\varepsilon \mathbf{e}_2,$$
$$\tilde{\zeta}_\varepsilon \boldsymbol{\phi}_{\zeta\mathbf{z}} \mathbf{e}_1 = \zeta_\varepsilon(\Re\, z\mathbf{e}_1 + \Im\, z\mathbf{e}_2).$$

Thus

$$\tilde{\zeta}_\varepsilon \Re z \mathbf{e}_1 + (\tilde{\zeta}_\varepsilon \Im\, z + \tilde{T}_\varepsilon \omega)\mathbf{e}_2 = \mathbf{0} \tag{7.33}$$

at $\varepsilon = 0$, whence $\tilde{\zeta}_\varepsilon(0) = 0 = \tilde{T}_\varepsilon(0)$.

We differentiate (7.32) and evaluate the derivative at $\varepsilon = 0$:

$$\mathbf{0} = \mathbf{U}_{\varepsilon\varepsilon}(0)\mathbf{e}_1 + (\mathbf{U}(0) - \mathbf{I})\tilde{\mathbf{z}}_{\varepsilon\varepsilon\varepsilon}(0) + 2\tilde{T}_{\varepsilon\varepsilon}(0)\boldsymbol{\phi}_{t\mathbf{z}}\mathbf{e}_1 + 2\tilde{\zeta}_{\varepsilon\varepsilon}(0)\boldsymbol{\phi}_{\zeta\mathbf{z}}\mathbf{e}_1. \tag{7.34}$$

We apply $\mathbf{P}_\mathrm{c}$ and take the inner product with $\mathbf{e}_1$. Since $\mathbf{P}_\mathrm{c}\tilde{\mathbf{z}}_{\varepsilon\varepsilon\varepsilon}(\varepsilon) = \mathbf{0}$ by construction, it follows that $\mathbf{P}_\mathrm{c}\mathbf{U}(0)\tilde{\mathbf{z}}_{\varepsilon\varepsilon\varepsilon}(0) = \mathbf{0}$, and thus we arrive at the crucial identity

$$\langle \mathbf{e}_1, \mathbf{P}_\mathrm{c}\mathbf{U}_{\varepsilon\varepsilon}(0)\mathbf{e}_1\rangle + 2\Re\, z\tilde{\zeta}_{\varepsilon\varepsilon}(0) = 0. \tag{7.35}$$

We need our special coordinates for this identity. In general coordinates it looks much more complicated.

The Floquet multipliers of the periodic solutions parametrized by ε are defined in terms of Poincaré sections. A useful Poincaré section for us is the half space $\mathfrak{P} = \mathcal{X}_\mathrm{h} \times \{\mathbf{z}_\mathrm{c} : \mathbf{z}_\mathrm{c} = \lambda\mathbf{e}_1 \text{ for } \lambda > 0\}$. Let $\mathbf{Q}(\varepsilon)$ be the projection onto $\mathcal{X}_\mathrm{h} \times \langle \mathbf{e}_1\rangle$ along $\varepsilon^{-1}\boldsymbol{\phi}_t(\tilde{T}(\varepsilon), \tilde{\mathbf{z}}(\varepsilon), \tilde{\zeta}(\varepsilon))$ where $\langle \mathbf{e}_1\rangle$ denotes the span of $\mathbf{e}_1$. The Floquet multipliers are the elements of the spectrum of $\mathbf{V}(\varepsilon) := \mathbf{Q}(\varepsilon)\boldsymbol{\phi}_\mathbf{z}(\tilde{T}(\varepsilon), \tilde{\mathbf{z}}(\varepsilon), \tilde{\zeta}(\varepsilon))$ where $\mathbf{V}(\varepsilon)$ is understood as an operator on $\mathcal{X}_\mathrm{h} \times \langle \mathbf{e}_1\rangle$. The periodic solution is stable if the spectrum of $\mathbf{V}(\varepsilon)$ is contained in the open unit ball in the complex plane. It is unstable if a part of the spectrum is outside the closed unit ball. Clearly $\mathbf{Q}$ is the projection along $\mathbf{e}_2$ in the limit $\varepsilon \to 0$ and $\mathbf{Q}(0) = \mathbf{P}_\mathrm{h} + \mathbf{P}_{\mathbf{e}_1}\mathbf{P}_\mathrm{c}$ where $\mathbf{P}_{\mathbf{e}_1}$ denotes the projection along $\mathbf{e}_2$ to $\langle \mathbf{e}_1\rangle$ in $\mathcal{X}_1 = \mathbf{e}_1$, and 1 is a a simple isolated eigenvalue of $\mathbf{V}(0)$. $\mathbf{V}$ depends differentiably on ε. Thus there exists a smooth family of eigenvalues $\mu(\varepsilon)$ of $\mathbf{V}(\varepsilon)$ with $\mu(0) = 0$. The other part of the spectrum is contained in the open unit ball for small ε if this is true at $\varepsilon = 0$. The assertion about stability now follows from the formulas

$$\mu_\varepsilon(0) = 0, \qquad \mu_{\varepsilon\varepsilon}(0) = -\Re\, z\tilde{\zeta}_{\varepsilon\varepsilon}(0), \tag{7.36}$$

which we shall prove below.

There exists a parametrized family of eigenvectors $\mathbf{e}(\varepsilon)$, which we assume to be smooth and normalized by $\mathbf{P}_{\mathbf{e}_1}\mathbf{P}_\mathrm{c}\mathbf{e}(\varepsilon) = \mathbf{e}_1$. We evualuate the derivative of the identity

$$\mathbf{V}(\varepsilon)\mathbf{e}(\varepsilon) = \mu(\varepsilon)\mathbf{e}(\varepsilon) \tag{7.37}$$

at $\varepsilon = 0$ to get

$$\mathbf{V}_\varepsilon(0)\mathbf{e}_1 + \mathbf{V}(0)\mathbf{e}_\varepsilon(0) = \mu_\varepsilon(0)\mathbf{e}_1 + \mathbf{e}_\varepsilon(0). \tag{7.38}$$

Clearly

$$\varepsilon^{-1}\boldsymbol{\phi}_t(\tilde{T}(\varepsilon),\tilde{\mathbf{z}}(\varepsilon),\tilde{\zeta}(\varepsilon)) \to \alpha(0)/\pi e_2 \quad \text{as} \quad \varepsilon \to 0. \tag{7.39}$$

We shall later show that $\frac{d}{d\varepsilon}[\varepsilon^{-1}\boldsymbol{\phi}_t(\tilde{\mathbf{z}}(\varepsilon),\tilde{T}(\varepsilon),\tilde{\zeta}(\varepsilon))] \to \mathbf{0}$ as $\varepsilon \to 0$. Thus

$$\mathbf{Q}_\varepsilon(0) = \mathbf{O}, \tag{7.40}$$

and, since $\mathbf{U}_\varepsilon \mathbf{e}_1 = \mathbf{0}$,

$$\mu_\varepsilon(0)\mathbf{e}_1 = \mathbf{V}(0)\mathbf{e}_\varepsilon(0) - \mathbf{e}_\varepsilon(0). \tag{7.41}$$

Hence, because of the normalization, we get

$$\mathbf{V}(0)\mathbf{e}_\varepsilon(0) - \mathbf{e}_\varepsilon(0) = \mathbf{0}, \tag{7.42}$$

which implies that

$$\mathbf{e}_\varepsilon(0) = \mathbf{0} \tag{7.43}$$

and that $\mu_\varepsilon(0) = 0$. Let

$$\mathbf{j}(\varepsilon) := \frac{d^2}{d\varepsilon^2}\boldsymbol{\phi}_t(\tilde{T}(\varepsilon),\tilde{\mathbf{z}}(\varepsilon),\tilde{\zeta}(\varepsilon)) = \frac{d}{d\varepsilon}(\boldsymbol{\phi}_{t\mathbf{z}}\tilde{\mathbf{z}}_\varepsilon + \boldsymbol{\phi}_{tt}\tilde{T}_\varepsilon + \boldsymbol{\phi}_{t\zeta}\tilde{\zeta}_\varepsilon), \tag{7.44}$$

which we evaluate at $\varepsilon = 0$:

$$\mathbf{j}(0) = \frac{d}{d\varepsilon}(\boldsymbol{\phi}_{t\mathbf{z}}\mathbf{e}_1) = \boldsymbol{\phi}_{t\mathbf{zz}} : \mathbf{e}_1\mathbf{e}_1 = \mathbf{0} \tag{7.45}$$

because $\tilde{T}_\varepsilon(0) = 0$, $\tilde{\zeta}_\varepsilon(0) = 0$, $\boldsymbol{\phi}_{tt}(t,0,\zeta) = \mathbf{0}$, and $\boldsymbol{\phi}_{t\zeta}(t,0,\zeta)) = \mathbf{0}$. The last identity in (7.45) holds because

$$\boldsymbol{\phi}_{\mathbf{zz}}(t,\mathbf{0},\zeta) : \mathbf{e}_1\mathbf{e}_1 = \mathbf{0} \tag{7.46}$$

by the choice of local coordinates. This implies the statement of (7.39) by l'Hôpital's rule.

We differentiate (7.37) a second time and evaluate the derivative at $\varepsilon = 0$ to obtain

$$\mathbf{V}_{\varepsilon\varepsilon}\mathbf{e}_1 + \mathbf{V}\mathbf{e}_{\varepsilon\varepsilon} - \mathbf{e}_{\varepsilon\varepsilon} = \mu_{\varepsilon\varepsilon}\mathbf{e}_1 + \mathbf{e}_{\varepsilon\varepsilon} \qquad \text{at } \varepsilon = 0. \tag{7.47}$$

We apply $\mathbf{P}_{\mathrm{c}}$, take the scalar product with $\mathbf{e}_1$, and obtain

$$\begin{aligned} \mu_{\varepsilon\varepsilon} \quad &= \mathbf{P}_{\mathrm{c}}\mathbf{V}_{\varepsilon\varepsilon}\mathbf{e}_1 = \mathbf{P}_{\mathrm{c}}\mathbf{Q}_{\varepsilon\varepsilon}\mathbf{e}_1 + \mathbf{P}_{\mathbf{e}_1}\mathbf{P}_{\mathrm{c}}\mathbf{U}_{\varepsilon\varepsilon}\mathbf{e}_1 \\ &= \mathbf{P}_{\mathbf{e}_1}\mathbf{P}_{\mathrm{c}}\mathbf{U}_{\varepsilon\varepsilon}\mathbf{e}_1 = -\Re z\tilde{\zeta}_{\varepsilon\varepsilon}\mathbf{e}_1 \quad \text{at } \varepsilon = 0, \end{aligned} \tag{7.48}$$

where the second equality holds because of (7.39), the third equality holds because $\mathbf{Q}(\varepsilon)\mathbf{e}_1 = \mathbf{0}$ for all ε, and the fourth equality is a consequence of (7.35). ∎

8. Discussion. We have shown that under reasonable and weak assumptions, the initial-boundary-value problem (1.3) defines a smooth semiflow on a nonlinear manifold. Moreover, we reduced the question of determining whether a Hopf bifurcation occurs to determining how the disposition of eigenvalues depends on the parameter ζ; this has to be carried out on a case-by-case basis. This is done in [7] for Example 2.1. In general, the assumptions on the eigenvalues could be checked numerically.

We had to use a nonstandard approach because all the abstract approaches we are aware of require the space of admissible initial values to be a vector space. Even though we can transform the local semiflow to a semiflow on a linear space, we do not know how to apply the usual techniques. The difficulty is the following: In the abstract approach one tries to understand the problem as an ordinary differential equation

$$\dot{x} = G(x)$$

in a Banach space with an unbounded nonlinear operator G. Linear problems of this type are well understood, and using the theory for them one may tackle the nonlinear problem.

In our case, G is defined on a Banach manifold. Let $\boldsymbol{\psi}$ be a local coordinate map. Then the flow can be expressed in local coordinates by

$$(\mathbf{z}, t) \to \boldsymbol{\psi}^{-1}(\boldsymbol{\phi}(t, \boldsymbol{\psi}(\mathbf{z}))).$$

Formally, this can be differentiated to obtain the differential equation

$$\dot{\mathbf{z}} = (\boldsymbol{\psi}^{-1})' \circ F(\boldsymbol{\psi}(\mathbf{z}), t).$$

In our case $F : C^{2,\alpha} \to C^{\alpha}$. At this point we cannot use our local coordinates, which are not continuous in the topology of C^{α}. Thus we could not apply the approach of [10], [17], [21], [22], [23], or [25], and we reverted to the approach of Hopf [14] as in [16]. On the other hand, these papers treat problems in all of $\mathbb{R}^n$, or problems with Dirichlet boundary conditions or periodic boundary conditions, so that they could work in a linear space.

Our study reveals that the solutions of the equations of viscoelasticity have some properties that are very similar to those for parabolic equations (e.g., the linearizations define analytic semigroups) and have other properties that are very different (e.g., noncompactness of the flow).

We have restricted our attention to scalar problems primarily for simplicity of exposition. This is not at all essential. In connection with Example 2.3 we have discussed the extension to systems. Similarly, it is likely that our approach works as well for other problems. This is certainly true for parabolic problems, but it should be true, for example, for some integro-differential equations of other kinds of viscoelasticity. Our methods can handle a variety of generalizations, e.g., problems in which Ω_T is not cylindrical.

Acknowledgment. We are grateful to the Mathematical Institute at Oberwolfach for providing facilities that enabled us to collaborate face-to-face at a critical stage of our work.

REFERENCES

[1] S. Agmon, A. Douglis, and L. Nirenberg, *Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions* II, Comm. Pure Appl. Math., 17 (1964), pp. 35–92.

[2] H. Amann, *Dynamic theory of quasilinear parabolic equations* I, Nonlinear Anal., 12 (1988), pp. 895–919.

[3] H. Amann, *Dynamic theory of quasilinear parabolic equations* II, Differential Integral Equations, 3 (1990), pp. 13–75.

[4] H. Amann, *Dynamic theory of quasilinear parabolic equations* III: *Global existence*, Math. Z., 202 (1989), pp. 219–250.

[5] S. S. Antman, *Nonlinear Problems of Elasticity*, Springer-Verlag, New York, 1995.

[6] S. S. Antman and C. A. Kenney, *Large buckled states of nonlinearly elastic rods under torsion, thrust, and gravity*, Arch. Rational Mech. Anal., 76 (1981), pp. 289–338.

[7] S. S. Antman and H. Koch, *Self-sustained oscillations of nonlinearly viscoelastic layers*, SIAM J. Appl. Math., 60 (2000), pp. 1357–1387.

[8] J. Cooper and H. Koch, *On the spectrum of a hyperbolic evolution operator*, J. Funct. Anal., 133 (1995), pp. 301–328.

[9] M. G. Crandall and P. H. Rabinowitz, *The Hopf bifurcation theorem in infinite dimensions*, Arch. Rational Mech. Anal., 67 (1977), pp. 53–72.

[10] G. Da Prato and A. Lunardi, *Stability, instability and center manifold theorem for fully nonlinear autonomous parabolic equations in Banach space*, Arch. Rational Mech. Anal., 74 (1988), pp. 115–141.

[11] A. Friedman, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, NJ, 1964.

[12] J. Guckenheimer and P. J. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Springer-Verlag, New York, 1983.

[13] D. Henry, *Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes in Math. 840, Springer-Verlag, New York, 1981.

[14] E. Hopf, *Abzweigung einer periodischen Lösung von einer stationären Lösung eines Differentialsystems*, Ber. Math-Phys. Sächsische Akademie der Wissenschaften, Leipzig, 94 (1942), pp. 1–22 (English translation by L. N. Howard and N. Kopell in [J. E. Mardsen and M. McCracken, Hopf Bifuraction and Its Applications, Springer-Verlag, New York, 1976]).

[15] L. N. Howard and N. Kopell, *Editorial comments*, in J. E. Marsden and M. McCracken, Hopf Bifurcation and Its Applications, Springer-Verlag, New York, 1976, pp. 194–205.

[16] H. Koch, *On a Fully Nonlinear Mixed Parabolic Problem with Oblique Boundary Condition*, Preprint SFB 359, Heidelberg, 1995.

[17] H. Koch, *On center manifolds*, Nonlinear Anal., 28 (1997), pp. 1227–1248.

[18] O. A. Ladyženskaja, H. Ural'ceva, and V. A. Solonnikov, *Linear and Quasi-Linear Equations of Parabolic Type*, AMS, Providence, RI, 1968.

[19] A. Lunardi, *Analytic Semi-Groups and Optimal Regularity in Parabolic Problems*, Birkhäuser, Basel, 1995.

[20] J. E. Marsden and M. McCracken, *Hopf Bifurcation and Its Applications*, Springer-Verlag, New York, 1976.

[21] M. Potier-Ferry, *The linearization principle for the stability of solutions of quasilinear parabolic equations.* I, Arch. Rational Mech. Anal., 77 (1981), pp. 301–320.

[22] M. Potier-Ferry, *On the mathematical foundations of elastic stability theory.* I, Arch. Rational Mech. Anal., 78 (1982), pp. 55–72.

[23] M. Renardy, *A centre manifold theorem for hyperbolic PDEs*, Proc. Roy. Soc. Edinburgh Sect. A, 122 (1992), pp. 363–377.

[24] V. A. Solonnikov, *On boundary-value problems for linear parabolic systems of differential equations of general form*, Proc. Steklov Inst. Math., 83 (1965), pp. 1–162 (English translation Amer. Math. Soc., Providence, RI, 1967, pp. 1–184.)

[25] C.-Y. Xu and J. E. Marsden, *Asymptotic stability of equilibria of nonlinear semiflows with applications to rotating viscoelastic rods, Part* 1, Topol. Methods Nonlinear Anal., 7 (1996), pp. 271–297.

SIAM J. MATH. ANAL.
Vol. 32, No. 2, pp. 385–402

STRONG ASYMPTOTICS OF ORTHONORMAL POLYNOMIALS WITH THE AID OF GREEN'S FUNCTION*

FRANZ PEHERSTORFER† AND ROBERT STEINBAUER†

Abstract. We give explicit asymptotic representations as well as ratio asymptotics of the orthogonal polynomials with asymptotically periodic reflection coefficients in terms of Green's function. All the limit relations will hold uniformly compact outside the support of the measure of orthogonality. Furthermore, with the help of harmonic measures we will characterize all those sets, i.e., supports of orthogonality measures, where orthogonal polynomials with asymptotically periodic reflection coefficients exist.

Key words. orthogonal polynomials, unit circle, arcs, asymptotics, asymptotically periodic reflection coefficients, Green's function, harmonic measures

AMS subject classifications. 33C45, 42C05

PII. S0036141098343045

1. Introduction.

Let $\{\Phi_n(z,\sigma) = \kappa_n z^n + \cdots\}_{n\in\mathbb{N}_0}$, $\kappa_n := \kappa_n(\sigma) > 0$, be a sequence of orthonormal polynomials on the unit circle with respect to the positive measure σ, i.e.,

$$\frac{1}{2\pi}\int_0^{2\pi} \Phi_n(e^{i\varphi},\sigma)\overline{\Phi_m(e^{i\varphi},\sigma)}\,d\sigma(\varphi) = \delta_{nm}.$$

All the measures appearing in this paper are understood to be probability measures, i.e., they are normalized by

$$\frac{1}{2\pi}\int_0^{2\pi} d\sigma(\varphi) = 1.$$

If the support of σ is an infinite set, then it is well known that the orthonormal polynomials are uniquely determined and can be completely described by the reflection coefficients

$$a_n := a_n(\sigma) =: \frac{\Phi_{n+1}(0,\sigma)}{\kappa_{n+1}}.$$

All these reflection coefficients are located in the open unit disk and generate the orthonormal polynomials iteratively by the recurrence relation

$$\sqrt{1-|a_n|^2}\,\Phi_{n+1}(z,\sigma) = z\Phi_n(z,\sigma) + a_n\Phi_n^*(z,\sigma), \quad n\in\mathbb{N}_0, \quad \Phi_0(z,\sigma) = 1, \tag{1.1}$$

*Received by the editors August 5, 1998; accepted for publication (in revised form) December 9, 1999; published electronically July 11, 2000. This work was supported by the Austrian Fonds zur Förderung der wissenschaftlichen Forschung, project P12985-TEC, and by a MAX-KADE postdoctoral fellowship, selected by the Österreichischen Akademie der Wissenschaften. This work was completed while the second author was visiting the Department of Mathematics, Ohio State University.

http://www.siam.org/journals/sima/32-2/34304.html

†Institut für Analysis und Numerik, Johannes Kepler Universität Linz, 4040 Linz-Auhof, Austria (franz.peherstorfer@jk.uni-linz.ac.at, robert.steinbauer@jk.uni-linz.ac.at).

where $\Phi_n^*(z,\sigma) := z^n\overline{\Phi_n(1/\bar{z},\sigma)}$ denotes the reversed polynomial. It is easy to see that the leading coefficients κ_n are given by

$$\kappa_n = \left(\prod_{j=0}^{n-1}(1-|a_j|^2)\right)^{-1/2}.$$

In this paper we study orthonormal polynomials whose reflection coefficients are asymptotically periodic. That is, there exists a periodic sequence $\{a_n^0\}_{n\in\mathbb{N}_0}$ with

$$a_{n+N}^0 = a_n^0, \quad n\in\mathbb{N}_0, \quad N\in\mathbb{N} \text{ fixed}, \quad |a_n^0|<1,$$

and

$$\lim_{\nu\to\infty} a_{\nu N+j} = a_j^0, \quad j=0,\dots,N-1. \tag{1.2}$$

We shall investigate the asymptotic behavior as well as ratio asymptotics of such orthonormal polynomials $\Phi_n(z,\sigma)$ as $n\to\infty$ outside the support of the orthogonality measure σ. So far, only comparative asymptotics for the perturbed and unperturbed orthonormal polynomials outside the support of the orthogonality measures were known (see [2, 23]). We shall derive explicit and strong limit representations in terms of Green's function, corresponding to the essential part of the support of the orthogonality measure, which consists of a finite union of arcs on the unit circle. For asymptotics on the support, see a forthcoming paper of the authors [24]. Furthermore, with the aid of harmonic measures we will characterize those unions of subarcs of the unit circle on which orthonormal polynomials with asymptotically periodic reflection coefficients live. By the way, let us recall the well-known fact (see, e.g., [25, section 11.5]) that polynomials orthogonal on a bounded set of the real line can be considered as special cases of polynomials orthogonal on the unit circle. Hence all the above results in particular give corresponding results for polynomials orthogonal on several intervals of the real line. For asymptotic representations of such polynomials and for related references, see [7] and [19]. Finally let us mention that the investigated polynomials play, apart from their use in numerical analysis, also an important role in problems of physics when the spectrum consists of several bands, for instance, in certain models of solid state physics [13], [28] and of nonlinear waves and solitons [16, 26, 4] but also in the field of complex iteration [3].

The proofs of all of our results are given at the end in section 5.

2. Some important preliminary notations and facts. Let us consider the periodic sequence of reflection coefficients $\{a_n^0\}_{n\in\mathbb{N}_0}$, $|a_n^0|<1$, and $a_{n+N}^0 = a_n^0$ for all $n\in\mathbb{N}_0$; $N\in\mathbb{N}$ fixed. The corresponding orthonormal polynomials are denoted by $\{\Phi_n(z,\sigma_0)\}_{n\in\mathbb{N}_0}$. It is well known that the polynomials of the second kind, defined by

$$\Psi_n(z,\sigma_0) := \frac{1}{2\pi}\int_0^{2\pi}\frac{e^{i\varphi}+z}{e^{i\varphi}-z}\left(\Phi_n(e^{i\varphi},\sigma_0)-\Phi_n(z,\sigma_0)\right)d\sigma_0(\varphi), \quad n\in\mathbb{N}, \tag{2.1}$$

and $\Psi_0(z,\sigma_0) := 1$ can be generated recursively in the same way as the orthonormal polynomials $\Phi_n(z,\sigma_0)$ (see (1.1)) by using the sequence of reflection coefficients $\{-a_n^0\}_{n\in\mathbb{N}_0}$ instead of $\{a_n^0\}_{n\in\mathbb{N}_0}$.

If we consider the m-shifted, $m\in\mathbb{N}$ fixed, sequence of reflection coefficients $\{a_{n+m}^0\}_{n\in\mathbb{N}_0}$ and if we proceed as in (1.1) and (2.1), respectively, we get the mth

associated orthonormal polynomials $\{\Phi_n^{[m]}(z,\sigma_0)\}_{n\in\mathbb{N}_0}$ and $\{\Psi_n^{[m]}(z,\sigma_0)\}_{n\in\mathbb{N}_0}$, respectively, which will be used in Corollary 3.5.

Now let us define the value

$$L := 2\left(\prod_{j=0}^{N-1}\left(1-|a_j^0|^2\right)\right)^{1/2} \tag{2.2}$$

and the monic polynomials

$$\begin{aligned}\mathcal{T}(z) &:= \frac{1}{2}\big(P_N(z,\sigma_0)+\Omega_N(z,\sigma_0)+P_N^*(z,\sigma_0)+\Omega_N^*(z,\sigma_0)\big) = z^N+\cdots,\\ R(z) &:= \mathcal{T}^2(z)-L^2z^N = z^{2N}+\cdots,\end{aligned} \tag{2.3}$$

where $\{P_n(z,\sigma_0)\}_{n\in\mathbb{N}_0}$ and $\{\Omega_n(z,\sigma_0)\}_{n\in\mathbb{N}_0}$ denote the monic orthogonal polynomials and the polynomials of the second kind, respectively, i.e.,

$$\begin{aligned}P_n(z,\sigma_0) &= \frac{\Phi_n(z,\sigma_0)}{\kappa_n(\sigma_0)} = \left(\prod_{j=0}^{n-1}(1-|a_j^0|^2)\right)^{1/2}\Phi_n(z,\sigma_0),\\ \Omega_n(z,\sigma_0) &= \left(\prod_{j=0}^{n-1}(1-|a_j^0|^2)\right)^{1/2}\Psi_n(z,\sigma_0).\end{aligned}$$

It can be shown that $\mathcal{T}$ and R have all their zeros on the unit circle; all the zeros of $\mathcal{T}$ are simple and those of R are at most double.

Remark. Let us note that for convenience of the reader we use a slightly different notation than in [22, 23]. Here, in contrast to [22, 23], R may have double zeros. So we can write

$$R(z) =: R^0(z)\mathcal{U}^2(z);$$

compare (4.1) below, where R^0 and $\mathcal{U}$ are self-reversed polynomials and where $\mathcal{U}$ vanishes exactly at the double zeros of R. Now R^0 corresponds to R from [22, 23].

We continue with the definition of the real set

$$E_l := \big\{\varphi\in[0,2\pi] : |\mathcal{T}(e^{i\varphi})|\le L\big\}.$$

From Geronimus's paper [10] one can derive that E_l consists of l, $l\le N$, disjoint intervals, while all the sets $\big\{\varphi\in[0,2\pi] : |\mathcal{T}(e^{i\varphi})|\le K\big\}$, with $0<K<L$, consist of exactly N intervals. That is, $l<N$ occurs if and only if there exist points ψ's such that $|\mathcal{T}(e^{i\varphi})|$ has a local extremum at ψ and $|\mathcal{T}(e^{i\psi})|=L$. In such a case, R has a double root at $e^{i\psi}$. It has also been shown by Geronimus [8, 9, 10] for the real case (see also [15]) that

$$\big(\operatorname{supp}(\sigma_0)\big)' = E_l$$

and $\operatorname{supp}(\sigma_0)\setminus E_l$ is a finite set. Let

$$\Gamma_{E_l} := \{e^{i\varphi} : \varphi\in E_l\}$$

be the corresponding arcs on the unit circle. Let us point out that all the zeros of the polynomial R are located on Γ_{E_l}, where each boundary point of Γ_{E_l} is a simple zero of R.

By (1.2) the $\Phi_n(z,\sigma)$'s can be considered as a compact perturbation of the "periodic" orthogonal polynomials $\Phi_n(z,\sigma_0)$. Hence, it is known that

$$\big(\mathrm{supp}(\sigma)\big)' = \big(\mathrm{supp}(\sigma_0)\big)' = \Gamma_{E_l}; \tag{2.4}$$

see [9] and [12, Theorem 3] for the case $N = 1$. (In fact, the proof given in [12] can easily be extended to the general case $N \in \mathbb{N}$.)

For the study of the asymptotic behavior of the orthonormal polynomials $\Phi_n(z,\sigma)$ outside the support of their orthogonality measure we exclude the case that all the a_j^0's are zero, which would imply that $\lim a_n = 0$; hence,

$$\sum_{j=0}^{N-1} |a_j^0| > 0. \tag{2.5}$$

If (2.5) is not fulfilled, one can apply the known theory dealing with orthogonal polynomials from Geronimus' class (i.e., $\sum |a_n| < \infty$) or from Szegö's class (i.e., $\sum |a_n|^2 < \infty$) or from Nevai's class (i.e., $a_n \to 0$).

Throughout this paper we will use the sequence of polynomials $\{\Phi_n(z,\sigma_0)\}_{n\in\mathbb{N}_0}$ with periodic reflection coefficients as a comparison system.

Let us start our considerations with the following settings:

$$\begin{aligned} y^{\pm}(z) &:= \frac{\mathcal{T}(z) \pm \sqrt{R(z)}}{L}, \qquad z \in \mathbb{C} \setminus \Gamma_{E_l}, \\ \Phi_n^{\pm}(z,\sigma_0) &:= \Phi_{n+N}(z,\sigma_0) - y^{\pm}(z)\Phi_n(z,\sigma_0), \\ \Phi_n^{*,\pm}(z,\sigma_0) &:= \Phi_{n+N}^*(z,\sigma_0) - y^{\pm}(z)\Phi_n^*(z,\sigma_0), \end{aligned}$$

where we take that branch of the square root $\sqrt{R}$ for which $\sqrt{R(0)} = 1$. By the above statements on the zeros of R, both functions $y^{\pm}$ are analytic on $\mathbb{C} \setminus \Gamma_{E_l}$ and the following estimates hold (cf. [23, Lemma 2.1]):

$$|y^+(z)| > 1 \quad \text{and} \quad |y^-(z)| < 1 \quad \text{on } \mathbb{C} \setminus \Gamma_{E_l}. \tag{2.6}$$

The functions $\{\Phi_n^{\pm}\}_{n\in\mathbb{N}_0}$ and $\{\Phi_n^{*,\pm}\}_{n\in\mathbb{N}_0}$ satisfy the same recurrence relation (1.1) as the orthonormal polynomials $\{\Phi_n(z,\sigma_0)\}_{n\in\mathbb{N}_0}$ and they will be important in the following.

Next, let us define two finite sets $\mathcal{N}$ and $\mathcal{N}^*$ by

$$\begin{aligned} \mathcal{N} &:= \big\{z \in \mathbb{C} \setminus \Gamma_{E_l} : \Phi_m^-(z,\sigma_0) = 0, m \in \{0,\dots,N-1\}\big\}, \\ \mathcal{N}^* &:= \big\{z \in \mathbb{C} \setminus \Gamma_{E_l} : \Phi_m^{*,-}(z,\sigma_0) = 0, m \in \{0,\dots,N-1\}\big\}. \end{aligned}$$

It will turn out that the points from $\mathcal{N}$ and $\mathcal{N}^*$, respectively, are exceptional in $\mathbb{C} \setminus \Gamma_{E_l}$ in the sense that the orthogonal polynomials $\Phi_n(z,\sigma_0)$ behave asymptotically differently at these points.

Since $\mathcal{N}$ and $\mathcal{N}^*$ will play a crucial role in what follows, let us describe basic facts of these sets. As a first remarkable property we note that by

$$\big(\mathcal{N} \cap \{|\,z\,| = 1\}\big) = \big(\mathcal{N}^* \cap \{|\,z\,| = 1\}\big) =: \{e^{i\xi_1},\dots,e^{i\xi_p}\} \tag{2.7}$$

one can describe exactly the mass-points of σ_0, i.e., $\sigma_0(\{\xi_1\}),\dots,\sigma_0(\{\xi_p\}) > 0$; see [23, Remark 2.2]. Furthermore,

$$\mathcal{N} \subset \{|z| \le 1\}, \qquad \mathcal{N}^* \subset \{|z| \ge 1\},$$

and

$$\mathcal{N}^* = \{1/\bar{z} : z \in \mathcal{N} \setminus \{0\}\}.$$

It will be interesting for the following to remark that the set $\mathcal{N}$ may indeed contain points from the interior of the unit circle. The simplest way to see this is to consider the sequence of reflection coefficients $\{a_n(\sigma_0)\} = \{0, a, 0, a, \dots\}$, $a \neq 0$; i.e., $N = 2$. Then

$$\Phi_{2k+1}(0, \sigma_0) = P_{2k+1}(0, \sigma_0) = 0 \quad \text{for all } k \in \mathbb{N}_0.$$

Hence,

$$\Phi_1^-(0, \sigma_0) = \Phi_3(0, \sigma_0) - y^-(0)\Phi_1(0, \sigma_0) = 0$$

and $0 \in \mathcal{N}$. But also $\mathcal{N} \cap \{0 < |z| < 1\} \neq \emptyset$ is possible as examples show.

On the other hand, $\mathcal{N} \cap \{|z| < 1\} \neq \emptyset$ is only possible for periods $N \geq 2$. For the period-1-case the set $\mathcal{N}$ is either empty, if σ_0 has no mass-point, or consists exactly of the point $e^{i\xi_1}$, where ξ_1 is the only mass-point of σ_0. This can easily be seen from the relation [23, (2.41)]

$$\mathcal{N} \subseteq \{z \in \mathbb{C} : P_1(z, \sigma_0) = \Omega_1(z, \sigma_0)\} \cup \{e^{i\xi_1}\},$$

where the set on the right-hand side is either equal to $\{e^{i\xi_1}\}$ if there is a ξ_1 or empty otherwise, since $P_1(z, \sigma_0) = z + a_0(\sigma_0)$, $\Omega_1(z, \sigma_0) = z - a_0(\sigma_0)$, and $a_0(\sigma_0) \neq 0$.

Finally, let us introduce the function

$$Y(z) := \left(\frac{\mathcal{T}(z) + \sqrt{R(z)}}{L}\right)^{1/N} = \sqrt[N]{y^+(z)}, \quad z \in \mathbb{C} \setminus \Gamma_{E_l}. \tag{2.8}$$

Note that by this definition the function Y is multiple-valued if $N > 1$, but both Y^N and $|Y|$ are single-valued on $\mathbb{C} \setminus \Gamma_{E_l}$. By (2.6) Y maps $\mathbb{C} \setminus \Gamma_{E_l}$ onto $\{|\,z\,| > 1\}$, the exterior of the unit disk.[1] Furthermore,

$$Y(z) = \mathcal{O}(z) \quad \text{as } z \text{ tends to infinity.}$$

On the arcs we have

$$|Y(z)| = \lim_{r \to 1} |Y(rz)| = 1, \quad z \in \Gamma_{E_l}.$$

Hence $\ln|Y(z)|$ is the real Green's function of $\mathbb{C} \setminus \Gamma_{E_l}$ with pole at infinity (for the definition of Green's function, see, e.g., [18, section 5.2] or [27, section 9.7]; compare also the beginning of section 4). Further, by

$$\big(\mathcal{T}(z) - \sqrt{R(z)}\big)\big(\mathcal{T}(z) + \sqrt{R(z)}\big) = L^2 z^N$$

and by the self-reversed property $\mathcal{T} = \mathcal{T}^*$ and $R = R^*$, one obtains

$$\lim_{z \to \infty} \frac{\mathcal{T}(z) - \sqrt{R(z)}}{L} = \lim_{\substack{y \to 0 \\ y = 1/\bar{z}}} \overline{\left[\frac{\mathcal{T}(y) - \sqrt{R(y)}}{L y^N}\right]} = \frac{L}{2}.$$

[1] Recall that we always assume (2.5) to hold true. Otherwise Γ_{E_l} would be the entire unit circle, $\mathbb{C} \setminus \Gamma_{E_l}$ would be disconnected, and the following statements would hold true only for $\{|z| > 1\}$ and not for $\mathbb{C} \setminus \Gamma_{E_l}$.

From this relation we derive

$$\lim_{z\to\infty}\frac{Y(z)}{z} = \sqrt[N]{2/L} =: \frac{1}{C(\Gamma_{E_l})},$$

i.e.,

$$C(\Gamma_{E_l}) = \sqrt[N]{L/2} = \left(\prod_{j=0}^{N-1}\left(1-|a_j^0|^2\right)\right)^{1/2N}$$

is the capacity of the set Γ_{E_l}.

3. Asymptotics of the orthonormal polynomials. In this section we study asymptotics of orthonormal polynomials with asymptotically periodic reflection coefficients. This section is organized as follows: First we will describe the asymptotic behavior of the "periodic" orthonormal polynomials $\{\Phi_n(z,\sigma_0)\}$ on the entire complex plane (see Proposition 3.1 and the remark thereafter). Next we shall state our asymptotic results concerning the polynomials $\{\Phi_n(z,\sigma)\}$ and give some additional remarks. All the proofs will be given in section 5.

Throughout this section we assume that the reflection coefficients $\{a_n\}$ converge sufficiently fast towards the periodic sequence $\{a_n^0\}$. To be precise, we suppose that

$$\sum_{n=0}^{\infty}|a_n - a_n^0| < \infty. \tag{3.1}$$

Let us recall the estimates

$$|Y(z)| > 1 \quad \text{on } \mathbb{C}\setminus\Gamma_{E_l} \tag{3.2}$$

and

$$\left|\frac{z}{Y(z)}\right| < 1 \quad \text{on } \mathbb{C}\setminus\Gamma_{E_l}, \tag{3.3}$$

which follow from (2.6), (2.8), and the relation $\mathcal{T}^2(z) - R(z) = L^2 z^N$.

With the inequalities (3.2) and (3.3) in mind the asymptotic behavior of the "periodic" orthonormal polynomials $\Phi_n(z,\sigma_0)$ outside the arcs Γ_{E_l} can be completely described with the aid of the function $Y(z)$ as follows.

PROPOSITION 3.1. *Let* (2.5) *be fulfilled. Then the following asymptotics hold for every* $m \in \{0,\dots,N-1\}$:

$$\begin{aligned}\lim_{\nu\to\infty}\left[\Phi_{m+\nu N}(z,\sigma_0) - Y^{\nu N}(z)\frac{L}{2}\frac{\Phi_m^-(z,\sigma_0)}{\sqrt{R(z)}}\right] = 0,\\ \lim_{\nu\to\infty}\left[\Phi^*_{m+\nu N}(z,\sigma_0) - Y^{\nu N}(z)\frac{L}{2}\frac{\Phi_m^{*,-}(z,\sigma_0)}{\sqrt{R(z)}}\right] = 0\end{aligned} \tag{3.4}$$

uniformly on compact subsets of $\mathbb{C}\setminus\Gamma_{E_l}$. *Furthermore,*

$$\Phi_{m+\nu N}(z,\sigma_0) = \left(\frac{z}{Y(z)}\right)^{\nu N}\frac{(-L)}{2}\frac{\Phi_m^+(z,\sigma_0)}{\sqrt{R(z)}} \underset{\nu\to\infty}{\longrightarrow} 0 \qquad \textit{for } z\in\mathcal{N},$$

$$\Phi^*_{m+\nu N}(z,\sigma_0) = \left(\frac{z}{Y(z)}\right)^{\nu N}\frac{(-L)}{2}\frac{\Phi_m^{*,+}(z,\sigma_0)}{\sqrt{R(z)}} \underset{\nu\to\infty}{\longrightarrow} 0 \qquad \textit{for } z\in\mathcal{N}^*.$$

We will also need the following lemma.

LEMMA 3.2. *Let us define the functions,* $n = \nu N + m$,

$$\mathcal{G}_n(z,\sigma_0) := -\frac{L\big(y^-(z)\big)^\nu}{z^n A(z)}\,\Phi_m^+(z,\sigma_0),$$
$$\mathcal{H}_n(z,\sigma_0) := -\frac{L\big(y^-(z)\big)^\nu}{z^{n+1} A(z)}\,\Phi_m^{*,+}(z,\sigma_0),$$

where the polynomial A *is given by* $A(z) = P_N^*(z,\sigma_0) - P_N(z,\sigma_0)$. *Then, under the assumptions* (2.5) *and* (3.1),

$$\lim_{n\to\infty}\big(\Phi_n^*(z,\sigma)\mathcal{G}_n(z,\sigma_0) - z\Phi_n(z,\sigma)\mathcal{H}_n(z,\sigma_0)\big) =: \Delta(z) \tag{3.5}$$

exists uniformly compact on $\mathbb{C}\setminus\big(\Gamma_{E_l}\cup\{e^{i\xi_1},\dots,e^{i\xi_p}\}\big)$; *notation as in* (2.7). *Furthermore, the function* Δ *is analytic and has no zeros outside the unit circle.*

From Proposition 3.1 we can derive strong asymptotics for the asymptotically periodic orthonormal polynomials $\Phi_n(z,\sigma)$ as follows.

THEOREM 3.3. *Suppose that* (2.5) *and* (3.1) *are fulfilled. Further, let the functions* Y *and* $\Delta(z)$ *be given in* (2.8) *and* (3.5), *respectively. Then for all* $k\in\mathbb{N}_0$ *the relation*

$$\lim_{\nu\to\infty}\frac{\Phi_{m+\nu N}^{(k)}(z,\sigma)}{(\nu N)_k Y^{\nu N-k}(z)} = \Delta(z)\big(Y'(z)\big)^k\frac{L}{4}\,\frac{\Phi_m^-(z,\sigma_0)}{\sqrt{R(z)}} \tag{3.6}$$

holds uniformly on compact subsets of $\mathbb{C}\setminus\big(\Gamma_{E_l}\cup\mathcal{N}\cup\mathcal{M}_k\big)$, *and the relation*

$$\lim_{\nu\to\infty}\frac{\big(\Phi_{m+\nu N}^*(z,\sigma)\big)^{(k)}}{(\nu N)_k Y^{\nu N-k}(z)} = \Delta(z)\big(Y'(z)\big)^k\frac{L}{4}\,\frac{\Phi_m^{*,-}(z,\sigma_0)}{\sqrt{R(z)}} \tag{3.7}$$

holds uniformly on compact subsets of $\mathbb{C}\setminus\big(\Gamma_{E_l}\cup\mathcal{N}^*\cup\mathcal{M}_k\big)$, *where*

$$\mathcal{M}_k = \begin{cases}\emptyset & \text{if } k=0,\\ \{Y'=0\} & \text{if } k\ge 1.\end{cases}$$

Here, $f^{(k)}(z)$ *denotes the* k*th derivative of the function* f *with respect to* z *and* $(n)_k := n(n-1)\times\cdots\times(n-k+1)$ *(and* $(n)_0 := 1$*).*

The next theorem gives important comparative asymptotics and will also be needed for the proof of Theorem 3.3.

THEOREM 3.4. *Suppose that* (2.5) *and* (3.1) *are fulfilled and let* $k\ge 0$ *be an arbitrary integer. Then, with the same notation as in Theorem* 3.3, *we have*

$$\lim_{n\to\infty}\frac{\Phi_n^{(k)}(z,\sigma)}{\Phi_n^{(k)}(z,\sigma_0)} = \frac{1}{2}\Delta(z) \tag{3.8}$$

uniformly on compact subsets of $\mathbb{C}\setminus\big(\Gamma_{E_l}\cup\mathcal{N}\cup\mathcal{M}_k\big)$ *and*

$$\lim_{n\to\infty}\frac{\big(\Phi_n^*(z,\sigma)\big)^{(k)}}{\big(\Phi_n^*(z,\sigma_0)\big)^{(k)}} = \frac{1}{2}\Delta(z) \tag{3.9}$$

uniformly on compact subsets of $\mathbb{C} \setminus \big(\Gamma_{E_l} \cup \mathcal{N}^* \cup \mathcal{M}_k\big)$.

For the following corollaries let us consider the zeros of Δ, i.e., the set $\{\Delta = 0\}$, where the function Δ is given in (3.5). Then

$$\{\Delta = 0\} \subset \{|\,z\,| = 1\} \quad \text{and} \quad \{e^{i\varphi} : \varphi \in \operatorname{supp}(\sigma) \setminus E_l\} \subseteq \{\Delta = 0\}$$

(see [23, Remark 3.2]).

COROLLARY 3.5. *Let* k, j *be any nonnegative integers and suppose that* (2.5) *and* (3.1) *are satisfied. Then*

$$\lim_{\nu\to\infty} \frac{\Phi_{(\nu N+k)+j}(z,\sigma)}{\Phi_{\nu N+k}(z,\sigma)} = \frac{1}{2}\left(\frac{\Phi_k^{*,-}(z,\sigma_0)}{\Phi_k^{-}(z,\sigma_0)}\, D_j^k(z) + C_j^k(z)\right) \tag{3.10}$$

holds uniformly on compact subsets of $\mathbb{C} \setminus \big(\Gamma_{E_l} \cup \mathcal{N} \cup \{\Delta = 0\}\big)$. *Here the polynomials* C_j^k *and* D_j^k *are given by*

$$\begin{aligned} C_j^k(z) &:= \Phi_j^{[k]}(z,\sigma_0) + \Psi_j^{[k]}(z,\sigma_0),\\ D_j^k(z) &:= \Phi_j^{[k]}(z,\sigma_0) - \Psi_j^{[k]}(z,\sigma_0); \end{aligned}$$

recall the definition of the k*th associated polynomials from section* 2.

Corollary 3.5 shows that the sequence $\{\Phi_{n+j}(z,\sigma)/\Phi_n(z,\sigma)\}_{n\in\mathbb{N}_0}$ contains N convergent subsequences with N different limit-functions. If we take $j = N$, then $\Phi_{n+N}(z,\sigma)/\Phi_n(z,\sigma)$ is convergent as n tends to infinity as follows.

COROLLARY 3.6. *Let* (2.5) *and* (3.1) *be fulfilled. Then there holds*

$$\lim_{n\to\infty} \frac{\Phi_{n+N}(z,\sigma)}{\Phi_n(z,\sigma)} = \frac{\mathcal{T}(z) + \sqrt{R(z)}}{L} = Y^N(z) \tag{3.11}$$

uniformly on compact subsets of $\mathbb{C} \setminus \big(\Gamma_{E_l} \cup \mathcal{N} \cup \{\Delta = 0\}\big)$.

At the end of this section let us state some additional remarks.

Remark. As we learned quite recently, Barrios and López [2] have shown that the limit relation (3.11) holds true even under weaker assumptions. For the case $N = 1$ compare also a recent result of Bello and López [1, Theorem 1].

Remark. It is obvious that from Theorem 3.3 one immediately can derive strong asymptotics for the reproducing kernel-function

$$K_n(z,\xi;\sigma) := \sum_{k=0}^{n} \Phi_k(z,\sigma)\overline{\Phi_k(\xi,\sigma)}$$

outside the set $\Gamma_{E_l} \cup \mathcal{N} \cup \mathcal{N}^*$ by making use of the Christoffel–Darboux formula (cf. [25, Theorem 11.4.2])

$$K_n(z,\xi;\sigma) = \frac{\Phi_{n+1}^*(z,\sigma)\overline{\Phi_{n+1}^*(\xi,\sigma)} - \Phi_{n+1}(z,\sigma)\overline{\Phi_{n+1}(\xi,\sigma)}}{1 - z\bar{\xi}}.$$

Remark. Let us point out that Theorem 3.4, and consequently also Theorem 3.3, does not hold true in general at the points from $\mathcal{N}$ and $\mathcal{N}^*$, respectively. An easy way to see this is to consider the reflection coefficients

$$\begin{aligned} \{a_n(\sigma_0)\} &:= \{0, a, 0, a, 0, \dots\},\\ \{a_n(\sigma)\} &:= \{b_0, a, b_1, a, b_2, \dots\}, \end{aligned}$$

where $b_n \to 0$ sufficiently fast such that (3.1) is fulfilled. Suppose further that all the b_n's are different from zero. Then $0 \in \mathcal{N}$ and

$$\Phi_{2m+1}(0,\sigma) = b_m \neq 0 \qquad \text{while} \qquad \Phi_{2m+1}(0,\sigma_0) = 0 \qquad \text{for all } m \in \mathbb{N}_0.$$

Hence, the quotient $\Phi_n(z,\sigma)/\Phi_n(z,\sigma_0)$ does not exist at $z = 0$ for all odd n, and consequently,

$$\frac{\Phi_n(0,\sigma)}{\Phi_n(0,\sigma_0)} \not\to \frac{1}{2}\Delta(0).$$

Since $\{Y' = 0\} \neq \emptyset$ is possible, one also has to stay away from this set if $k > 0$.

Remark. Even though (3.6) does not hold true on the set $\mathcal{N}$ in general, we can say at least that

$$\frac{\Phi_n^{(k)}(z,\sigma)}{n^k Y^n(z)} = \mathcal{O}(1) \quad \text{on } \mathcal{U}(z) \tag{3.12}$$

for every neighborhood $\mathcal{U}(z)$ of $z \in \mathcal{N}$ with $\mathcal{U}(z) \cap \Gamma_{E_l} = \emptyset$; $k \in \mathbb{N}_0$ fixed. This can easily be derived from the maximum principle since the functions at the left-hand side of (3.12) are analytic on $\mathcal{U}(z)$ and uniformly bounded on $\partial\mathcal{U}(z)$. An analogous estimate as in (3.12) also holds true for the reversed polynomials $\big(\Phi_n^*(z,\sigma)\big)^{(k)}$ on $\mathcal{N}^*$.

4. Characterization of the arcs. In this section we will characterize those subsets of the unit circle, where orthogonal polynomials with asymptotically periodic reflection coefficients exist. Such subsets will be unions of finitely many arcs which we again will denote by Γ_{E_l}. To be precise, we shall describe those arcs Γ_{E_l}, respectively, intervals E_l, which coincide with the accumulation points of the support of a measure σ, i.e., $(\mathrm{supp}(\sigma))' = E_l$, whose corresponding reflection coefficients behave asymptotically periodic.

Additional to the notations in the previous sections, let

$$E_l =: \bigcup_{j=1}^{l} [\varphi_{2j-1}, \varphi_{2j}], \quad \varphi_i \neq \varphi_k \text{ for } i \neq k.$$

Of course, there is a one-to-one relation between the arcs $\Gamma_{E_l} := \bigcup_{j=1}^{l} \{e^{i\varphi} : \varphi \in [\varphi_{2j-1}, \varphi_{2j}]\}$ and the self-reversed polynomials, which vanish exactly at the boundary points of the arcs,

$$R^0(z) := \rho \prod_{j=1}^{2l} (z - e^{i\varphi_j}), \qquad \rho = (-1)^l \exp\left\{ -\frac{i}{2} \sum_{j=1}^{2l} \varphi_j \right\}.$$

Let us point out that the polynomial R^0 is of degree $2l$ while the polynomial R, defined in (2.3), was of degree $2N \geq 2l$, where N denoted the length of the period of the asymptotically periodic reflection coefficients under consideration. Furthermore, R had $N - l$ additional double zeros in the interior of the arcs. Hence, we can write

$$R(z) =: R^0(z)\mathcal{U}^2(z), \tag{4.1}$$

where $\mathcal{U}$ is again a self-reversed polynomial of degree $N - l$ whose zeros are all simple and are located in the interior of the arcs.

To state the next theorem we need to recall the following definitions.

By $g(z) := g(z,\infty)$ we denote the (real) Green's function for the set $\mathbb{C} \setminus \Gamma_{E_l}$ with pole at ∞. This function is uniquely determined by the following properties:

(1) $g(z)$ is harmonic on $\mathbb{C} \setminus \Gamma_{E_l}$,
(2) $g(z) - \ln|z|$ is harmonic near ∞,
(3) $\lim_{z \to e^{i\varphi}} g(z) = 0$ for all $\varphi \in E_l$
(compare, e.g., [18, section 5.2] or [27, section 9.7]).

The harmonic measure $\omega(\Gamma_j) := \omega(\Gamma_j, \infty)$ at ∞ of the arcs

$$\Gamma_j := \Gamma_{[\varphi_{2j-1}, \varphi_{2j}]} = \{e^{i\varphi} : \varphi \in [\varphi_{2j-1}, \varphi_{2j}]\}, \quad j = 1, \ldots, l,$$

is given by

$$\omega(\Gamma_j) = \lim_{\gamma_j} \frac{1}{2\pi} \int_{\gamma_j} \frac{\partial}{\partial n_\zeta} g(\zeta)\, |d\zeta| =: \frac{1}{2\pi} \int_{\partial \Gamma_j} \frac{\partial}{\partial n_\zeta} g(\zeta)\, |d\zeta|, \quad j = 1, \ldots, l. \tag{4.2}$$

Here, γ_j means a closed Jordan curve which encycles Γ_j counterclockwise and where the limit over γ_j is understood to approximate the arc Γ_j, i.e., the integration of the second integral in (4.2) is performed twice, once from $e^{i\varphi_{2j-1}}$ to $e^{i\varphi_{2j}}$, where $(\partial/\partial n_\zeta) g(\zeta)$ is approached from outside the unit disk, and once from $e^{i\varphi_{2j}}$ to $e^{i\varphi_{2j-1}}$, where $(\partial/\partial n_\zeta) g(\zeta)$ is approached from inside the unit disk. Furthermore, n_ζ denotes the unit normal derivative at $\zeta \in \Gamma_j$ and $|d\zeta|$ is the differential of arc length on Γ_j (see, e.g., [17, Chapter I.10, p. 38], [29, section 4, p. 140]).

Let us give a more convenient representation of the harmonic measures:

LEMMA 4.1. *Let $\Gamma_{E_l} = \bigcup_{j=1}^{l} \Gamma_j$, where $\Gamma_j := \{e^{i\varphi} : \varphi \in [\varphi_{2j-1}, \varphi_{2j}]\}$, be a union of arcs on the unit circle. Then there exists a uniquely determined polynomial $S_l = S_l^*$ of degree l, normalized by $i\, S_l(0) = \sqrt{R^0(0)}$, which satisfies*

$$\int_{\varphi_{2j}}^{\varphi_{2j+1}} \frac{S_l(e^{i\varphi})}{\sqrt{R^0(e^{i\varphi})}}\, d\varphi = 0 \quad \text{for } j = 1, \ldots, l-1. \tag{4.3}$$

Furthermore, the harmonic measures $\omega(\Gamma_j)$, $j = 1, \ldots, l$, are of the form

$$\omega(\Gamma_j) = \frac{1}{2\pi} \int_{\varphi_{2j-1}}^{\varphi_{2j}} \left| \frac{S_l(e^{i\varphi})}{\sqrt{R^0(e^{i\varphi})}} \right| d\varphi. \tag{4.4}$$

Now, the following characterization holds.

THEOREM 4.2. *The following statements are equivalent:*

(i) *There exists a measure σ with corresponding asymptotically N-periodic reflection coefficients*[2] *and $(\mathrm{supp}(\sigma))' = E_l$, $l \le N$.*

(ii) *There exist self-reversed polynomials $\mathcal{T}(z) = \mathcal{T}^*(z) = z^N + \cdots$ of degree l, $l \le N$, and $\mathcal{U}(z) = \mathcal{U}^*(z) = \beta z^{N-l} + \cdots$ of degree $N - l$ such that the quadratic polynomial equation*

$$\mathcal{T}^2(z) - R^0(z)\mathcal{U}^2(z) = L^2 z^N \qquad \text{with} \quad L > 0 \tag{4.5}$$

holds.

(iii) *There exists a self-reversed polynomial $S_l = S_l^*$ of degree l with $i\, S_l(0) = \sqrt{R^0(0)} = \beta$, and positive integers k_j, $j = 1, \ldots, l$, satisfying $\sum_{j=1}^{l} k_j = N$, such that*

$$\int_{\varphi_{2j}}^{\varphi_{2j+1}} \frac{S_l(e^{i\varphi})}{\sqrt{R^0(e^{i\varphi})}}\, d\varphi = 0 \quad \text{for } j = 1, \ldots, l-1 \tag{4.3}$$

[2] Note that we assume only that $\lim_{\nu \to \infty} a_{\nu N + j} = a_j^0$, $j = 0, \ldots, N-1$, and do not suppose the strong condition (3.1).

and

$$\int_{\varphi_{2j-1}}^{\varphi_{2j}} \left| \frac{S_l(e^{i\varphi})}{\sqrt{R^0(e^{i\varphi})}} \right| d\varphi = \frac{2k_j\pi}{N} \quad \textit{for } j = 1, \dots, l. \tag{4.6}$$

(iv) *There exist positive integers* k_j, $j = 1, \dots, l$ *satisfying* $\sum_{j=1}^{l} k_j = N$, *such that*

$$\omega(\Gamma_j) = \frac{k_j}{N} \quad \textit{for } j = 1, \dots, l.$$

Remark. (a) For alternative characterizations for the existence of self-reversed polynomials $\mathcal{T}$ and $\mathcal{U}$ which satisfy a quadratic equation of the form (4.5), i.e., for the existence of a "periodic" measure σ_0, see also [22, section 3].

(b) For the case of several real intervals the counterpart of Theorem 4.2 can be found in [19, pp. 191–194].

5. Proofs. *Proof of Proposition* 3.1 *and Lemma* 3.2. The limit-relations in Proposition 3.1 are rewritings of the authors' results [23, Theorem 2.1 and Remark 2.1] and Lemma 3.2 is [23, Theorem 3.1]. □

Since we will use Theorem 3.4 for the proof of Theorem 3.3, we start with the following.

Proof of Theorem 3.4. For $k = 0$ this is Theorem 3.2 in [23]. It suffices to prove our theorem only for $k = 1$ since the general statements then follow immediately by an induction argument. Let us write

$$\frac{\Phi_n'(z,\sigma)}{\Phi_n'(z,\sigma_0)} = \frac{\Phi_n(z,\sigma_0)}{\Phi_n'(z,\sigma_0)} \left(\frac{\Phi_n(z,\sigma)}{\Phi_n(z,\sigma_0)} \right)' + \frac{\Phi_n(z,\sigma)}{\Phi_n(z,\sigma_0)}. \tag{5.1}$$

Since the theorem is true for $k = 0$, we know that

$$\lim_{n\to\infty} \frac{\Phi_n(z,\sigma)}{\Phi_n(z,\sigma_0)} = \frac{1}{2}\Delta(z) \tag{5.2}$$

and

$$\lim_{n\to\infty} \left(\frac{\Phi_n(z,\sigma)}{\Phi_n(z,\sigma_0)} \right)' = \frac{1}{2}\Delta'(z) \tag{5.3}$$

uniformly on compact subsets of $\mathbb{C} \setminus \big(\Gamma_{E_l} \cup \mathcal{N}\big)$. Note that Δ' exists everywhere on $\mathbb{C} \setminus \big(\Gamma_{E_l} \cup \mathcal{N}\big)$ and is analytic there, since Δ is analytic. Next we show that

$$\lim_{n\to\infty} \frac{\Phi_n(z,\sigma_0)}{\Phi_n'(z,\sigma_0)} = 0 \qquad \text{uniformly compact on } \mathbb{C} \setminus \big(\Gamma_{E_l} \cup \mathcal{N} \cup \{Y' = 0\}\big). \tag{5.4}$$

Together with (5.1) to (5.3) this will prove (3.8) for $k = 1$.

To see (5.4) let $m \in \{0, \dots, N-1\}$ be arbitrary but fixed. From (3.4) we obtain

$$\lim_{\nu\to\infty} \left(\Phi_{m+\nu N}'(z,\sigma_0) - \frac{L}{2} \left(Y^{\nu N}(z) \frac{\Phi_m^-(z,\sigma_0)}{\sqrt{R(z)}} \right)' \right) = 0 \tag{5.5}$$

uniformly compact on $\mathbb{C} \setminus \big(\Gamma_{E_l} \cup \mathcal{N}\big)$. But

$$\left(Y^{\nu N}(z) \frac{\Phi_m^-(z,\sigma_0)}{\sqrt{R(z)}} \right)' = (\nu N) Y^{\nu N-1}(z) Y'(z) \frac{\Phi_m^-(z,\sigma_0)}{\sqrt{R(z)}} + Y^{\nu N}(z) \left(\frac{\Phi_m^-(z,\sigma_0)}{\sqrt{R(z)}} \right)'.$$

The derivative $\big(\Phi_m^-(z,\sigma_0)/\sqrt{R(z)}\big)'$ exists everywhere on $\mathbb{C}\setminus\Gamma_{E_l}$. Now, (5.5) and (3.4) give

$$\lim_{\nu\to\infty}\frac{\Phi'_{m+\nu N}(z,\sigma_0)}{\nu N\Phi_{m+\nu N}(z,\sigma_0)}=\frac{Y'(z)}{Y(z)}$$

uniformly on compact subsets of $\mathbb{C}\setminus\big(\Gamma_{E_l}\cup\mathcal{N}\big)$. Since Y'/Y exists on $\mathbb{C}\setminus\Gamma_{E_l}$, is different from zero on $\mathbb{C}\setminus(\Gamma_{E_l}\cup\{Y'=0\})$, and is independent of m, (5.4) follows. This finishes the proof of (3.8).

In exactly the same way one also shows (3.9). □

Proof of Theorem 3.3. For the proof of Theorem 3.3 we have to put together the results of Proposition 3.1 and Theorem 3.4. Therefore, let us write

$$\frac{\Phi^{(k)}_{m+\nu N}(z,\sigma)}{(\nu N)_k Y^{\nu N-k}(z)}=\frac{\Phi^{(k)}_{m+\nu N}(z,\sigma)}{\Phi^{(k)}_{m+\nu N}(z,\sigma_0)}\cdot\frac{\Phi^{(k)}_{m+\nu N}(z,\sigma_0)}{(\nu N)_k Y^{\nu N-k}(z)}. \tag{5.6}$$

Since we know that the first factor at the right-hand side tends to $\Delta/2$ by Theorem 3.4, we have to study only the second factor. By (3.4) and the fact that $(\nu N)_k Y^{\nu N-k}\to\infty$ we have

$$\begin{aligned}\lim_{\nu\to\infty}\frac{\Phi^{(k)}_{m+\nu N}(z,\sigma_0)}{(\nu N)_k Y^{\nu N-k}(z)}&=\frac{L}{2}\lim_{\nu\to\infty}\frac{\big(Y^{\nu N}(z)\Phi_m^-(z,\sigma_0)/\sqrt{R(z)}\big)^{(k)}}{(\nu N)_k Y^{\nu N-k}(z)}\\&=\frac{L}{2}\big(Y'(z)\big)^k\ \frac{\Phi_m^-(z,\sigma_0)}{\sqrt{R(z)}}.\end{aligned} \tag{5.7}$$

The first part of the theorem follows now from (5.6) and (5.7). The asymptotics of $(\Phi_n^*(z,\sigma))^{(k)}$ are shown in the same way. □

Proof of Corollary 3.5. From the identity

$$2\Phi_{n+j}(z,\sigma)=\big(\Phi_n(z,\sigma)+\Phi_n^*(z,\sigma)\big)\Phi_j^{[n]}(z,\sigma)+\big(\Phi_n(z,\sigma)-\Phi_n^*(z,\sigma)\big)\Psi_j^{[n]}(z,\sigma)$$

(see [20, Corollary 3.1]) we obtain

$$\begin{aligned}&\frac{2\Phi_{(\nu N+k)+j}(z,\sigma)}{\Phi_{\nu N+k}(z,\sigma)}\\&\quad=\left(1+\frac{\Phi^*_{\nu N+k}(z,\sigma)}{\Phi_{\nu N+k}(z,\sigma)}\right)\Phi_j^{[\nu N+k]}(z,\sigma)+\left(1-\frac{\Phi^*_{\nu N+k}(z,\sigma)}{\Phi_{\nu N+k}(z,\sigma)}\right)\Psi_j^{[\nu N+k]}(z,\sigma).\end{aligned} \tag{5.8}$$

It is easy to see that

$$\begin{aligned}\lim_{\nu\to\infty}\Phi_j^{[\nu N+k]}(z,\sigma)&=\Phi_j^{[k]}(z,\sigma_0),\\\lim_{\nu\to\infty}\Psi_j^{[\nu N+k]}(z,\sigma)&=\Psi_j^{[k]}(z,\sigma_0)\end{aligned} \tag{5.9}$$

uniformly on each compact set, since all the involved polynomials are of fixed degree j. From Theorem 3.3 we obtain that

$$\lim_{\nu\to\infty}\frac{\Phi^*_{\nu N+k}(z,\sigma)}{\Phi_{\nu N+k}(z,\sigma)}=\frac{\Phi_k^{*,-}(z,\sigma_0)}{\Phi_k^-(z,\sigma_0)} \tag{5.10}$$

holds uniformly compact on $\mathbb{C} \setminus (\Gamma_{E_l} \cup \mathcal{N} \cup \mathcal{N}^* \cup \{\Delta = 0\})$. But all the functions in (5.10) are uniformly bounded around points from $\mathcal{N}^* \cap \{|z| > 1\}$; compare the last remark before section 4. Hence, Vitali's theorem shows that (5.10) holds uniformly on compact subsets of $\mathbb{C} \setminus (\Gamma_{E_l} \cup \mathcal{N} \cup \{\Delta = 0\})$. Together with (5.8) and (5.9) this finishes the proof. □

Proof of Corollary 3.6. Since the "periodic" orthonormal polynomials $\Phi_n(z, \sigma_0)$ satisfy the limit relation (3.10) as well, we obtain from Corollary 3.5 that

$$\lim_{n\to\infty} \left(\frac{\Phi_{n+N}(z,\sigma)}{\Phi_n(z,\sigma)} - \frac{\Phi_{n+N}(z,\sigma_0)}{\Phi_n(z,\sigma_0)} \right) = 0$$

uniformly on compact subsets of $\mathbb{C} \setminus \big(\Gamma_{E_l} \cup \mathcal{N} \cup \{\Delta = 0\}\big)$. Hence, it suffices to consider only the ratio $\Phi_{n+N}(z, \sigma_0)/\Phi_n(z, \sigma_0)$.

From Corollary 2.1(a) in [23] we have

$$\lim_{n\to\infty} \Phi_n^+(z, \sigma_0) = 0$$

uniformly on compact subsets of $\mathbb{C} \setminus \Gamma_{E_l}$. But

$$\Phi_n^+(z, \sigma_0) = \Phi_n(z, \sigma_0) \left(\frac{\Phi_{n+N}(z, \sigma_0)}{\Phi_n(z, \sigma_0)} - y^+(z) \right)$$

and $\Phi_n(z, \sigma_0) \to \infty$ uniformly on compact subsets of $\mathbb{C} \setminus (\Gamma_{E_l} \cup \mathcal{N})$ by Proposition 3.1. This gives the assertion. □

Proof of Lemma 4.1. The proof of the existence of the polynomial S_l is very similar to the one given in [29, pp. 225–226]. Recall that S_l has complex coefficients, and, hence, (4.3) can be considered as a linear system of $l - 1$ equations for $l - 1$ unknowns. If $l = 1$, the only condition is $i\, S_l(0) = \sqrt{R^0(0)}$.

Let g denote the real Green's function for the set $\mathbb{C} \setminus \Gamma_{E_l}$ with pole at ∞ and let $\tilde{g}$ be its (up to an additive constant) harmonic conjugate. By the complex Green's function G we understand the function

$$G(z) := g(z) + i\tilde{g}(z).$$

Now let us consider $g(\zeta) = g(x, y)$ with $\zeta = x + iy$. For $\zeta = e^{i\varphi} \in \Gamma_{E_l}$ we have

$$\begin{aligned}
\zeta G'(\zeta+) &= \zeta\big(g_x(\zeta+) + i\tilde{g}_x(\zeta+)\big) = \zeta\big(g_x(\zeta+) - ig_y(\zeta+)\big) \\
&= \big(g_x(\zeta+)\cos\varphi + g_y(\zeta+)\sin\varphi\big) - i\big(g_y(\zeta+)\cos\varphi - g_x(\zeta+)\sin\varphi\big) \\
&= \frac{d}{d_{n_\zeta}} g(\zeta+) - i\frac{d}{d_{t_\zeta}} g(\zeta+) \\
&= \frac{d}{d_{n_\zeta}} g(\zeta+),
\end{aligned}$$

where $g(\zeta+)$ means the limit of $g(y)$ as $y \to \zeta$ from outside the unit disk and where $(d/d_{t_\zeta})g(\zeta+)$ denotes the tangential derivative of g, which is zero since g vanishes identically on the arcs Γ_{E_l}. If we approach the arcs from the interior of the unit disk (now the unit normal is $-\zeta$) we get in the same way as above

$$\zeta G'(\zeta-) = -\frac{d}{d_{n_\zeta}} g(\zeta-).$$

Hence, the harmonic measures $\omega(\Gamma_j)$ from (4.2) are of the form

$$\omega(\Gamma_j) = \frac{1}{2\pi} \int_{\Gamma_j} \zeta\big(G'(\zeta+) - G'(\zeta-)\big)\, |d\zeta|.$$

In order to prove (4.4) we have to show that the derivative of the complex Green's function (which is single-valued) satisfies

$$\zeta\big(G'(\zeta+) - G'(\zeta-)\big) = \left| \frac{S_l(e^{i\varphi})}{\sqrt{R^0(e^{i\varphi})}} \right|, \quad \zeta = e^{i\varphi}, \tag{5.11}$$

since $|d\zeta| = d\varphi$.[3] Let us give an explicit representation of G. We claim that

$$G(z) = \frac{1}{2} \int_{z_1}^{z} \frac{1}{\zeta} \left(1 - \frac{iS_l(\zeta)}{\sqrt{R^0(\zeta)}} \right) d\zeta, \qquad z_1 := e^{i\varphi_1}, \tag{5.12}$$

where the integration is performed along a path in the complex plane cut along Γ_{E_l}. Indeed, G is analytic on $\mathbb{C} \setminus \Gamma_{E_l}$ with

$$G(z) = \log z + \mathcal{O}(1) \quad \text{as } z \to \infty,$$

since the integrand in (5.12) behaves as $2/\zeta$ for large ζ; compare (5.26). Finally, we have to show that

$$\lim_{\zeta \to z} G(\zeta) \text{ is purely imaginary for } z \in \Gamma_{E_l}. \tag{5.13}$$

Let us start to point out two facts: First, with $z_k := e^{i\varphi_k}$,

$$\operatorname{Re} \left\{ \int_{z_{2j}}^{z_{2j+1}} \frac{1}{\zeta} \left(1 - \frac{iS_l(\zeta)}{\sqrt{R^0(\zeta)}} \right) d\zeta \right\} = 0, \quad j = 1, \ldots, l-1,$$

where the path of integration is on the unit circle. This follows from (4.3) and $d\zeta/\zeta = i\, d\varphi$ for $\zeta = e^{i\varphi}$. Second, $iS_l(\zeta)/\sqrt{R^0(\zeta)}$ is real for $\zeta \in \Gamma_{E_l}$ and, hence, the integrand $\big[\big(1 - iS_l(\zeta)/\sqrt{R^0(\zeta)}\big)/\zeta\big]\, d\zeta$ in (5.12) is purely imaginary on Γ_{E_l}.

Now recall that the integral in (5.12) is independent of the path of integration as long as it does not cross the arcs Γ_{E_l}, because the integrand is analytic on $\mathbb{C} \setminus \Gamma_{E_l}$. From the two facts above, property (5.13) follows, and so G is indeed the complex Green's function on $\mathbb{C} \setminus \Gamma_{E_l}$ with pole at infinity. Now it is easy to see that

$$G'(z) = \frac{1}{2z} \left(1 - \frac{iS_l(z)}{\sqrt{R^0(z)}} \right),$$

and (5.11) follows, since $\sqrt{R(\zeta+)} = -\sqrt{R(\zeta-)}$ and $iS_l(\zeta)/\sqrt{R^0(\zeta)}$ is real and positive on Γ_{E_l}. This completes the proof. □

Proof of Theorem 4.2. First of all, let us point out that every measure σ with asymptotically periodic reflection coefficients can be considered as a compact perturbation of the measure σ_0 with periodic reflection coefficients and, therefore, the accumulation points of $\operatorname{supp}(\sigma)$ and $\operatorname{supp}(\sigma_0)$ coincide; compare (2.4). Hence, it suffices

[3] In case an equation of the form (4.5) holds, we have $G(z) = \log Y(z)$, where Y is defined as in (2.8), and $G'(z) = Y'(z)/Y(z)$. Then, by straightforward calculation, using (5.21), (5.22), and (4.5), property (5.11) follows.

to show that the parts (ii), (iii), and (iv) of the theorem are equivalent with the fact that there exists a "periodic" measure σ_0 on the arcs Γ_{E_l}, respectively, intervals E_l.

The equivalence of (i) and (ii) has been shown by the authors in [22, Theorem 4.4]. Let us now prove the equivalence of (ii) and (iii) by showing both implications.

(ii) *implies* (iii). Let us define the real trigonometric polynomials

$$\begin{aligned}\mathcal{R}^0(\varphi) &:= e^{-il\varphi}R^0(e^{i\varphi}),\\ \tau(\varphi) &:= e^{-i(N/2)\varphi}\mathcal{T}(e^{i\varphi}),\\ u(\varphi) &:= e^{-i((N-l)/2)\varphi}\mathcal{U}(e^{i\varphi}),\end{aligned}$$

which satisfy, as a consequence of (4.5),

$$\tau^2(\varphi) - \mathcal{R}^0(\varphi)u^2(\varphi) = L^2. \tag{5.14}$$

By the geometry of (5.14) we can write the derivative of τ in the form

$$\tau'(\varphi) = \frac{N}{2}u(\varphi)s_l(\varphi), \tag{5.15}$$

where s_l is a real trigonometric polynomial of degree $l/2$, and consequently

$$\mathcal{R}^0(\varphi)\big[\tau'(\varphi)\big]^2 = \frac{N^2}{4}\big(\tau^2(\varphi) - L^2\big)s_l^2(\varphi). \tag{5.16}$$

For $\varphi \in (\varphi_{2j}, \varphi_{2j+1})$, $j = 1, \dots, l-1$, we have

$$\mathcal{R}^0(\varphi) > 0, \qquad \text{hence,} \qquad \left|\frac{\tau(\varphi)}{L}\right| > 1,$$

where the second inequality follows from (5.14). It is not difficult to see that the above differential equation (5.16) has for $\varphi \in (\varphi_{2j}, \varphi_{2j+1})$ the solution

$$\frac{\tau(\varphi)}{L} = \pm\cosh\left(\frac{N}{2}\int_{\varphi_{2j}}^{\varphi}\frac{s_l(\psi)}{\sqrt{\mathcal{R}^0(\psi)}}\,d\psi\right). \tag{5.17}$$

Now recall that $\tau(\varphi_{2j}) = \tau(\varphi_{2j+1}) = \pm L$. Hence, if we set

$$S_l(z) := z^{l/2}s_l(\varphi), \quad z = e^{i\varphi}, \tag{5.18}$$

property (4.3) follows.

Next let $[\psi_1, \psi_2] \subseteq [\varphi_{2j-1}, \varphi_{2j}]$, $j = 1, \dots, l$, be the maximal subinterval, where τ is monotone. Then, again by (5.14)

$$\tau(\psi_1) = -\tau(\psi_2) = \pm L \quad \text{and} \quad \left|\frac{\tau(\varphi)}{L}\right| \le 1 \quad \text{on } [\psi_1, \psi_2].$$

Further,

$$\frac{\tau(\varphi)}{L} = \pm\cos\left(\frac{N}{2}\int_{\psi_1}^{\varphi}\frac{|s_l(\psi)|}{\sqrt{|\mathcal{R}^0(\psi)|}}\,d\psi\right) \tag{5.19}$$

solves the differential equation (5.16); and from $\tau(\psi_1)/L = -\tau(\psi_2)/L = \pm 1$ we obtain that

$$\frac{N}{2}\int_{\psi_1}^{\psi_2}\frac{|s_l(\psi)|}{\sqrt{|\mathcal{R}^0(\psi)|}}\,d\psi = \pi.$$

This gives (4.6).

It remains to show that the polynomial S_l, defined in (5.18), satisfies

$$S_l = S_l^* \quad \text{and} \quad i\,S_l(0) = \sqrt{R^0(0)}.$$

The first property is an immediate consequence of (5.18). For the second one we need the following considerations: We have

$$\tau'(\varphi) = \frac{d}{d\varphi}\big(e^{-i(N/2)\varphi}\mathcal{T}(e^{i\varphi})\big) = ie^{-i(N/2)\varphi}\left(e^{i\varphi}\mathcal{T}'(e^{i\varphi}) - \frac{N}{2}\mathcal{T}(e^{i\varphi})\right),$$

and multiplying this equation by $e^{i(N/2)\varphi}$ gives

$$e^{i(N/2)\varphi}\tau'(\varphi) = \widetilde{\mathcal{T}}(e^{i\varphi}), \tag{5.20}$$

where the polynomial $\widetilde{\mathcal{T}}$ is defined by

$$\widetilde{\mathcal{T}}(z) := iz\,\mathcal{T}'(z) - \frac{iN}{2}\,\mathcal{T}(z) = \frac{iN}{2}z^N + \cdots. \tag{5.21}$$

By (5.20) the polynomial $\widetilde{\mathcal{T}}$ is self-reversed, i.e., $\widetilde{\mathcal{T}} = \widetilde{\mathcal{T}}^*$, and from (5.15), (5.18), and (5.20) we obtain that

$$\widetilde{\mathcal{T}}(z) = \frac{N}{2}\,\mathcal{U}(z)S_l(z). \tag{5.22}$$

Comparing the leading coefficients gives

$$S_l(z) = \frac{i}{\beta}z^l + \cdots - \frac{i}{\overline{\beta}}. \tag{5.23}$$

From (4.1) we see that $|\beta| = 1$ and thus $S_l(0) = -i/\overline{\beta} = -i\beta = -i\sqrt{R^0(0)}$.

(iii) *implies* (ii). Motivated by (5.17), we define

$$\tau(z) := L\cosh\left(\frac{N}{2}\int_{\varphi_1}^{z}\frac{s_l(\xi)}{\sqrt{\mathcal{R}^0(\xi)}}\,d\xi\right), \tag{5.24}$$

where the path of integration is outside E_l and where the functions $s_l(z)$ and $\mathcal{R}^0(z)$, $z \in \mathbb{C}$, are the extensions of the real trigonometric polynomials

$$\begin{aligned} s_l(\varphi) &:= e^{-i(l/2)\varphi}S_l(e^{i\varphi}),\\ \mathcal{R}^0(\varphi) &:= e^{-il\varphi}R^0(e^{i\varphi}), \end{aligned} \qquad \varphi \in \mathbb{R}.$$

Obviously, τ is analytic on the cut complex plane $\mathbb{C} \setminus E_l$. A little bit of further investigation, using the well-known relation $\cosh(iz) = \cos(z)$, shows that τ is continuous on $\mathbb{R}$. Hence, τ is analytic on the entire complex plane. Next we will show that $\tau(\varphi)$, $\varphi \in \mathbb{R}$, is a real trigonometric polynomial of degree $N/2$. One way to do this is to prove that for every $y \in \mathbb{C}$

$$\mathcal{T}(y) := y^{N/2}\tau(-i\ln y) \tag{5.25}$$

defines a self-reversed algebraic polynomial of degree N, since then

$$\tau(\varphi) = e^{-i(N/2)\varphi}\mathcal{T}(e^{i\varphi}).$$

First we show that $\mathcal{T}$ is an entire function, i.e., analytic on the entire complex plane: From (4.3) and (4.6) one obtains that

$$\tau(z+2\pi) = (-1)^N \tau(z) \text{ for all } z \in \mathbb{C}.$$

Hence, it doesn't matter which branch of the logarithm we take in (5.25) and we end up with an analytic function on $\mathbb{C} \setminus \{0\}$. But $\mathcal{T}$ is also continuous, and thus analytic, at $y = 0$, which can be seen as follows: By our assumption on S_l we have

$$\frac{S_l(0)}{\sqrt{R^0(0)}} = -i \qquad \text{and} \qquad \lim_{\zeta\to\infty} \frac{S_l(\zeta)}{\sqrt{R^0(\zeta)}} = i \tag{5.26}$$

and hence

$$\int_{\varphi_1}^{-i\ln y} \frac{S_l(e^{i\xi})}{\sqrt{R^0(e^{i\xi})}}\, d\xi = \mathcal{O}(\ln y) \quad \text{as } y \to 0 \text{ or } \infty.$$

Now it is not difficult to see that

$$\mathcal{T}(y) = y^{N/2} L \cosh\left(\frac{N}{2} \int_{\varphi_1}^{-i\ln y} \frac{S_l(e^{i\xi})}{\sqrt{R^0(e^{i\xi})}}\, d\xi \right) = \mathcal{O}(1) \quad \text{as } y \to 0, \tag{5.27}$$

and $\mathcal{T}$ is indeed an entire function. In a way very similar as in (5.27) one obtains

$$\frac{\mathcal{T}(y)}{y^N} = \mathcal{O}(1) \qquad \text{as} \quad y \to \infty, \tag{5.28}$$

and by Liouville's theorem $\mathcal{T}$ has to be a polynomial of degree N. Using definition (5.25) we see that $\mathcal{T}$ is also self-reversed.

Summing up, we have shown that τ from (5.24) is for $z = \varphi \in \mathbb{R}$ a real trigonometric polynomial. Again by the relation $\cosh(iz) = \cos(z)$ we obtain the estimates

$$\begin{aligned} |\tau(\varphi)| &\geq L \quad \text{on } \mathbb{R} \setminus E_l, \\ |\tau(\varphi)| &\leq L \quad \text{on } E_l. \end{aligned}$$

As an immediate consequence, τ has $N + l$ extremal points on E_l. By [22, Corollary 3.2(c)] this is sufficient for the existence of self-reversed polynomials $\mathcal{T}$ and $\mathcal{U}$ satisfying (4.5); in fact, the polynomial $\mathcal{T}$ is given by (5.25), respectively, (5.27). This completes the proof of the equivalence of (ii) and (iii).

The remaining equivalence of (iii) and (iv) is just a consequence of Lemma 4.1. □

Acknowledgment. The authors thank Professor Paul Nevai of the Ohio State University for numerous conversations regarding the results of this paper. In particular, he suggested a generalization of [23, Theorem 3.2], which is included here as Theorem 3.4.

REFERENCES

[1] M. Bello and G. López, *Ratio and relative asymptotics of polynomials on an arc of the unit circle*, J. Approx. Theory, 92 (1998), pp. 216–244.

[2] D. Barrios Rolania and G. López, *Ratio asymptotics for polynomials on arcs of the unit circle*, Constr. Approx., 15 (1999), pp. 1–31.

[3] D. Bessis, *Orthogonal polynomials, Padé approximations and Julia sets*, in Orthogonal Polynomials: Theory and Practice, P. Nevai, ed., NATO, Adv. Sci. Inst. Ser. C Math. Phys. Sci. 294, Kluwer, Dordrecht, The Netherlands, 1990, pp. 55–97.
[4] P. Deift, T. Kriecherbauer, and S. Venakides, *Forced lattice vibrations: Part* I *and* II, Comm. Pure Appl. Math., 48 (1995), pp. 1187–1298.
[5] P. L. Duren, *Theory of H^p Spaces*, Pure and Applied Mathematics 38, Academic Press, New York, London, 1970.
[6] G. Freud, *Orthogonal Polynomials*, Pergamon, Oxford, New York, Toronto, 1971.
[7] J. S. Geronimo and W. Van Assche, *Orthogonal polynomials with asymptotically periodic recurrence coefficients*, J. Approx. Theory, 46 (1986), pp. 251–283.
[8] Ya. L. Geronimus, *On the character of the solutions of the moment problem in the case of a limit-periodic associated fraction*, Izv. Akad. Nauk SSSR Ser. Math., 5 (1941), pp. 203–210 (in Russian).
[9] Ya. L. Geronimus, *On polynomials orthogonal on the circle, on the trigonometric moment-problem and on allied Caratheodory and Schur functions*, C.R. (Doklady) Acad. Sci. URSS (N.S.), 39 (1943), pp. 291–295.
[10] Ya. L. Geronimus, *On polynomials orthogonal on a circle, on the trigonometric moment-problem, and on the associated functions of Caratheodory's and Schur's type*, Rec. Math. [Mat. Sbornik] N.S., 15 (57), (1944), pp. 99–130 (in Russian).
[11] Ya. L. Geronimus, *Polynomials orthogonal on a circle and their applications*, Amer. Math. Soc. Translation 1954, (1954), pp. 1–78.
[12] L. Golinskii, P. Nevai, and W. Van Assche, *Perturbation of orthogonal polynomials on an arc of the unit circle*, J. Approx. Theory, 83 (1995), pp. 392–422.
[13] R. Haydock, *The recursive solution of the Schrödinger equation*, Solid State Physics, 35 (1980), pp. 215–294.
[14] N. Levinson, *Simplified treatment of integrals of Cauchy type, the Hilbert problem and singular integral equations. Appendix: Poincaré–Bertrand formula*, SIAM Rev., 7 (1965), pp. 474–502.
[15] A. Máté, P. Nevai and W. Van Assche, *The supports of measures associated with orthogonal polynomials and the spectra of the related self-adjoint operators*, Rocky Mountain J. Math., 21 (1991), pp. 501–527.
[16] P. D. Miller, N. M. Ercolani, and I. M. Krichever, *Finite genus solutions to the Ablowitz-Ladik equations*, Comm. Pure Appl. Math., 48 (1995), pp. 1369–1440.
[17] Z. Nehari, *Conformal Mapping*, Dover, New York, 1952.
[18] E. M. Nikishin and V. N. Sorokin, *Rational Approximations and Orthogonality*, Transl. Math. Monogr. 92, Amer. Math. Soc., Providence, RI, 1991.
[19] F. Peherstorfer, *Orthogonal and extremal polynomials on several intervals*, J. Comput. Appl. Math., 48 (1993), pp. 187–205.
[20] F. Peherstorfer, *A special class of polynomials orthogonal on the unit circle including associated polynomials*, Constr. Approx., 12 (1996), pp. 161–186.
[21] F. Peherstorfer and R. Steinbauer, *Comparative asymptotics for perturbed orthogonal polynomials*, Trans. Amer. Math. Soc., 348 (1996), pp. 1459–1486.
[22] F. Peherstorfer and R. Steinbauer, *Orthogonal polynomials on arcs of the unit circle.* II. *Orthogonal polynomials with periodic reflection coefficients*, J. Approx. Theory, 87 (1996), pp. 60–102.
[23] F. Peherstorfer and R. Steinbauer, *Asymptotic behaviour of orthogonal polynomials on the unit circle with asymptotically periodic reflection coefficients*, J. Approx. Theory, 88 (1997), pp. 316–353.
[24] F. Peherstorfer and R. Steinbauer, *Asymptotic behaviour of orthogonal polynomials on the unit circle with asymptotically periodic reflection coefficients,* II. *Weak asymptotics*, J. Approx. Theory, to appear.
[25] G. Szegö, *Orthogonal Polynomials*, 4th ed., Amer. Math. Soc. Colloq. Publ., 23, AMS, Providence, RI, 1975.
[26] M. Toda, *Nonlinear Waves and Solitons*, Kluwer, Dordrecht, The Netherlands, 1989.
[27] L. V. Toralballa, *Theory of Functions*, Charles E. Merrill, Columbus, OH, 1963.
[28] D. G. Pettifor and D. L. Weaire, eds., *The Recursion Method and Its Applications*, Springer Ser. Solid-State Sci. 58, Springer-Verlag, Berlin, Heidelberg, New York, Tokyo, 1985.
[29] H. Widom, *Extremal polynomials associated with a spectrum of curves in the complex plane*, Adv. Math., 3 (1969), pp. 127–232.

SIAM J. MATH. ANAL.
Vol. 32, No. 2, pp. 403–419

A MONOTONICITY PROPERTY INVOLVING ${}_3F_2$ AND COMPARISONS OF THE CLASSICAL APPROXIMATIONS OF ELLIPTICAL ARC LENGTH*

ROGER W. BARNARD†, KENT PEARCE†, AND KENDALL C. RICHARDS‡

Abstract. Conditions are determined under which ${}_3F_2\left(-n,a,b;a+b+2,\varepsilon-n+1;1\right)$ is a monotone function of n satisfying $ab\cdot{}_3F_2\left(-n,a,b;a+b+2,\varepsilon-n+1;1\right)\geq ab\cdot{}_2F_1\left(a,b;a+b+2;1\right)$. Motivated by a conjecture of Vuorinen [*Proceedings of Special Functions and Differential Equations*, K. S. Rao, R. Jagannathan, G. Vanden Berghe, J. Van der Jeugt, eds., Allied Publishers, New Delhi, 1998], the corollary that ${}_3F_2(-n,-\frac{1}{2},-\frac{1}{2};1,\varepsilon-n+1;1)\geq\frac{4}{\pi}$, for $1>\epsilon\geq\frac{1}{4}$ and $n\geq 2$, is used to determine surprising hierarchical relationships among the 13 known historical approximations of the arc length of an ellipse. This complete list of inequalities compares the Maclaurin series coefficients of ${}_2F_1$ with the coefficients of each of the known approximations, for which maximum errors can then be established. These approximations range over four centuries from Kepler's in 1609 to Almkvist's in 1985 and include two from Ramanujan.

Key words. hypergeometric, approximations, elliptical arc length

AMS subject classifications. 33C, 41A

PII. S003614109935050X

1. Introduction.

Let $L(x,y)$ be the arc length of an ellipse with semiaxes of length x and y (with $x\geq y>0$) and let $\lambda\equiv\dfrac{x-y}{x+y}$. In 1742, Maclaurin [12] determined that

$$L(x,y)=\pi(x+y)\cdot{}_2F_1\left(-\frac{1}{2},-\frac{1}{2};1;\lambda^2\right), \tag{1}$$

where ${}_2F_1$ is the hypergeometric function defined by

$${}_2F_1\left(a,b;c;z\right)\equiv 1+\sum_{n=1}^{\infty}\frac{(a)_n(b)_n z^n}{(c)_n n!}$$

with the Appell (or Pochhammer) symbol $(a)_n\equiv a(a+1)\cdots(a+n-1)$ for $n\geq 1$ and $(a)_0\equiv 1$, $a\neq 0$. (For more background information, see [2], [14], [9], and the recent survey article [8] by the first author.) In [2], Almkvist and Berndt compiled and presented the list of the approximations in Table 1.1 for

$$G(\lambda)\equiv{}_2F_1\left(-\frac{1}{2},-\frac{1}{2};1;\lambda^2\right)=\frac{L(x,y)}{\pi(x+y)}.$$

These approximations and their historical and recent connections to the approximations of π can be found in the Borweins' book [10]. Another excellent source for historical and current studies of these topics is the book [5] by Anderson, Vamanamurthy, and Vuorinen.

*Received by the editors January 15, 1999; accepted for publication (in revised form) December 8, 1999; published electronically July 11, 2000.
http://www.siam.org/journals/sima/32-2/35050.html
†Department of Mathematics, Texas Tech University, Lubbock, TX 79409 (barnard@math.ttu.edu, pearce@math.ttu.edu)
‡Department of Mathematics, Southwestern University, Georgetown, TX 78626 (richards@southwestern.edu)

TABLE 1.1
Approximations of $G(\lambda) \equiv {}_2F_1(-\frac{1}{2}, -\frac{1}{2}; 1; \lambda^2)$ (see [2]).

Discoverer(s) and year of discovery	Approximation $A_p(\lambda)$	δ_p = first nonzero term in the Maclaurin series for $\Delta_p(\lambda) \equiv A_p(\lambda) - G(\lambda)$
Kepler, 1609	$A_1(\lambda) \equiv (1-\lambda^2)^{1/2}$	$\delta_1 = -\frac{3}{4}\lambda^2$
Euler, 1773	$A_2(\lambda) \equiv (1+\lambda^2)^{1/2}$	$\delta_2 = \frac{1}{4}\lambda^2$
Sipos, 1792 Ekwall, 1973	$A_3(\lambda) \equiv \dfrac{2}{1+\sqrt{1-\lambda^2}}$	$\delta_3 = \frac{7}{64}\lambda^4$
Peano, 1889	$A_4(\lambda) \equiv \frac{3}{2} - \frac{1}{2}(1-\lambda^2)^{1/2}$	$\delta_4 = \frac{3}{64}\lambda^4$
Muir, 1883	$A_5(\lambda) \equiv \left(\dfrac{(1+\lambda)^{3/2} + (1-\lambda)^{3/2}}{2}\right)^{2/3}$	$\delta_5 = -\frac{1}{64}\lambda^4$
Lindner, 1904-1920 Nyvoll, 1978	$A_6(\lambda) \equiv \left(1 + \dfrac{\lambda^2}{8}\right)^2$	$\delta_6 = -\frac{1}{2^8}\lambda^6$
Selmer, 1975	$A_7(\lambda) \equiv 1 + \dfrac{\lambda^2/4}{1-\lambda^2/16}$	$\delta_7 = -\frac{3}{2^{10}}\lambda^6$
Ramanujan, 1914 Fergestad, 1951	$A_8(\lambda) \equiv 3 - \sqrt{4-\lambda^2}$	$\delta_8 = -\frac{1}{2^9}\lambda^6$
Almkvist, 1978	$A_9(\lambda) \equiv 2\dfrac{\left(1+\sqrt{1-\lambda^2}\right)^2 + \lambda^2\sqrt{1-\lambda^2}}{\left(1+\sqrt{1-\lambda^2}\right)\left(1+\sqrt[4]{1-\lambda^2}\right)^2}$	$\delta_9 = \frac{15}{2^{14}}\lambda^8$
Bronshtein and Semendyayev, 1964 Selmer, 1975	$A_{10}(\lambda) \equiv \dfrac{64-3\lambda^4}{64-16\lambda^2}$	$\delta_{10} = -\frac{9}{2^{14}}\lambda^8$
Selmer, 1975	$A_{11}(\lambda) \equiv \frac{3}{2} + \dfrac{\lambda^2}{8} - \dfrac{1}{2}\left(1 - \dfrac{\lambda^2}{2}\right)^{1/2}$	$\delta_{11} = -\frac{5}{2^{14}}\lambda^8$
Jacobsen and Waadeland, 1985	$A_{12}(\lambda) \equiv \dfrac{256 - 48\lambda^2 - 21\lambda^4}{256 - 112\lambda^2 + 3\lambda^4}$	$\delta_{12} = -\frac{33}{2^{18}}\lambda^{10}$
Ramanujan, 1914	$A_{13}(\lambda) \equiv 1 + \dfrac{3\lambda^2}{10 + \sqrt{4-3\lambda^2}}$	$\delta_{13} = -\frac{3}{2^{17}}\lambda^{10}$

Recently, several inequalities between various mean values and the hypergeometric function were proved in [10], [15], and the dependence of the hypergeometric function ${}_2F_1(a, b; c; z)$ on its parameters was studied in [4], [6]. These results led to a conjecture of Vuorinen (see [16]) concerning Muir's approximation A_5. Vuorinen conjectured (see [16]) that

$$A_5(\lambda) \leq G(\lambda) \quad \text{for all } \lambda \in [0,1]. \tag{2}$$

That is, Vuorinen conjectured that A_5 is a *lower bound* for G. This conjecture was recently proved by the authors in [9] which has become the genesis of the present article. Moreover, the results here attest to the adage that a single conjecture may have many ramifications. Also, note that A_5 is one of the mean values studied in [15]. More approximations for hypergeometric functions in terms of such mean values are actively being sought. For example, let $\nu \in \mathbf{R}\backslash\{0\}$ and define

$$M_\nu(\lambda) \equiv \left[\frac{(1+\lambda)^\nu + (1-\lambda)^\nu}{2}\right]^{1/\nu}.$$

H. Alzer [3] originally made the following conjecture.

CONJECTURE. *The inequalities*

$$M_\alpha(\lambda) \le G(\lambda) \le M_\beta(\lambda) \quad \textit{hold for all } \lambda \in (0,1) \tag{3}$$

if and only if

$$\alpha \le 3/2 \quad \textit{and} \quad \beta \ge (\ln 2)/\left(\ln \frac{\pi}{2}\right) \approx 1.53.$$

As noted by Alzer [3], it follows from our results (see the set of inequalities in expression (4)) that (3) holds with $\alpha = 3/2$ and $\beta = 2$. Moreover, for a fixed λ, $M_\nu(\lambda)$ is an increasing function of ν. Thus it follows that (3) holds for all $\alpha \le 3/2$ and $\beta \ge 2$. It can be shown that $\alpha = 3/2$ is sharp.

2. Main results. In an earlier paper (see [9]), the authors were able to verify inequality (2) by working with the original version of Vuorinen's conjecture in terms of the eccentricity (see (5) and (6)). In this direction, a generating function argument (motivated by [7]) was used to obtain the following general result (which will also be applied in this paper to obtain Theorem 2.5).

THEOREM 2.1 (see [9]). *Suppose $a, b > 0$. Then for any ε satisfying $1 > \varepsilon \ge \frac{ab}{a+b+1}$, it follows that*

$${}_3F_2\left(-n, a, b; a+b+1, \varepsilon - n + 1; 1\right) \ge 0,$$

for all integers $n \ge 1$, where ${}_3F_2$ is the generalized hypergeometric function.

In light of the conjecture in (2), the following question naturally arises:

> *Which of the remaining approximations given in Table* 1.1 *are upper bounds or lower bounds for G?*

An attempt to compare an approximation A_p with G motivates an analysis of the term δ_p (the first nonzero term in the Maclaurin series representation for the error function $\Delta_p(\lambda) \equiv A_p(\lambda) - G(\lambda)$). What information does δ_p provide? Certainly the leading term can be viewed as a measure of accuracy of the given approximation, and the error function $\Delta_p(\lambda)$ will have the same sign as δ_p for *sufficiently small* λ. For example, $\delta_1 < 0$ and it follows directly that A_1 is a lower bound for G, as Kepler intended (see [2, p. 599]). In this case, the sign of δ_1 is indicative of the sign of $\Delta_1(\lambda)$ *for all* $\lambda \in [0,1]$. Almkvist and Berndt proved (see [2, p. 603]) that Ramanujan's first estimate A_8 is a lower bound for G by proving the significantly stronger result that the nonzero Maclaurin series coefficients of Δ_8 all have the same (negative) sign. A numerical investigation suggests that a similar trait might be shared by other approximations given in Table 1.1. In this article, it will be shown that all of the approximations given in Table 1.1 satisfy the following property:

> *The sign of the error function $\Delta_p(\lambda)$ coincides with the sign of the leading term δ_p for all $\lambda \in [0,1]$.*

Moreover, for all but two of the approximations, it will be established that the nonzero Maclaurin series coefficients of Δ_p all have the same sign as δ_p. (Only Euler's approximation and Muir's approximation fail to satisfy this condition.) As a consequence of the forthcoming results, each function $|\Delta_p|$ is a strictly increasing function of λ, for $p = 1, \ldots, 13$. Therefore, $0 = |\Delta_p(0)| < |\Delta_p(\lambda)| < |\Delta_p(1)|$ for all $\lambda \in (0,1)$. For example, the maximum error for Ramanujan's second estimate is $|\Delta_{13}(1)| = |\frac{14}{11} - \frac{4}{\pi}| \approx 0.000512$ and satisfies $|\Delta_{13}(1)| < |\Delta_p(1)|$ for $p = 1, \ldots, 12$. In this direction, we will prove the following three propositions.

PROPOSITION 2.2. *Let* $G(\lambda) \equiv \sum_{n=0}^{\infty} \alpha_n \lambda^{2n}$ *and* $A_p(\lambda) \equiv \sum_{n=0}^{\infty} \beta_n^{(p)} \lambda^{2n}$ *where* $\alpha_n \equiv (\frac{(-1/2)_n}{n!})^2$ *and each* A_p *is defined as in Table* 1.1. *Then*

$$\beta_n^{(12)} \leq \alpha_n \leq \beta_n^{(9)} \quad \text{for all integers } n \geq 0.$$

Therefore, the error functions $|\Delta_9|$ *and* $|\Delta_{12}|$ *are strictly increasing and*

$$A_{12}(\lambda) \leq G(\lambda) \leq A_9(\lambda) \quad \text{for all } \lambda \in [0,1].$$

PROPOSITION 2.3. *Let* $G(\lambda) \equiv \sum_{n=0}^{\infty} \alpha_n \lambda^{2n}$ *and* $A_p(\lambda) \equiv \sum_{n=0}^{\infty} \beta_n^{(p)} \lambda^{2n}$ *where* $\alpha_n \equiv (\frac{(-1/2)_n}{n!})^2$ *and each* A_p *is defined as in Table* 1.1. *Then*

$$\beta_n^{(1)} \leq \beta_n^{(6)} \leq \beta_n^{(7)} \leq \beta_n^{(8)} \leq \beta_n^{(10)} \leq \beta_n^{(11)} \leq \beta_n^{(13)} \leq \alpha_n \leq \beta_n^{(4)} \leq \beta_n^{(3)}$$

for all integers $n \geq 0$. *Therefore, the corresponding error functions* $|\Delta_p|$ *are strictly increasing and*

$$A_1(\lambda) \leq A_6(\lambda) \leq A_7(\lambda) \leq A_8(\lambda) \leq A_{10}(\lambda) \leq A_{11}(\lambda) \leq A_{13}(\lambda) \leq G(\lambda) \leq A_4(\lambda) \leq A_3(\lambda)$$

for all $\lambda \in [0,1]$.

The next proposition addresses the two remaining estimates: Euler's approximation A_2 and Muir's approximation A_5. The claim will be made that

$$A_5(\lambda) \equiv \left(\frac{(1+\lambda)^{3/2} + (1-\lambda)^{3/2}}{2} \right)^{2/3} \leq G(\lambda) \leq (1+\lambda^2)^{1/2} \equiv A_2(\lambda) \tag{4}$$

for all $\lambda \in [0,1]$. As we have noted, the nonzero Maclaurin series coefficients of Δ_2 and Δ_5 (as functions of λ) do not have constant sign. In order to verify the inequalities in (4), we make use of the known fact due to Landen and Ivory (e.g., see [2, p. 598]) that

$$G(\lambda) \equiv {}_2F_1\left(-\frac{1}{2}, -\frac{1}{2}; 1; \lambda^2\right) = \frac{2x}{x+y} \cdot {}_2F_1\left(\frac{1}{2}, -\frac{1}{2}; 1; \xi^2\right), \tag{5}$$

where $\lambda \equiv (x-y)/(x+y)$ and $\xi \equiv (1/x)\sqrt{x^2-y^2}$ is the eccentricity of the original ellipse (see (1)). Without loss of generality, assume that $1 = x \geq y \geq 0$. A change of variable from λ to ξ can be accomplished in (4) by using (5) and the substitutions $\lambda = (1-y)/(1+y)$ and $y = \sqrt{1-\xi^2}$. Multiplying through by $(1+y)/2$ and simplifying, we see that the inequalities in (4) are equivalent to

$$\left(\frac{1 + (1-\xi^2)^{3/4}}{2} \right)^{2/3} \leq {}_2F_1\left(\frac{1}{2}, -\frac{1}{2}; 1; \xi^2\right) \leq (1 - \xi^2/2)^{1/2} \tag{6}$$

for all $\xi \in [0,1]$. (The first inequality in (6) is the original version of Vuorinen's conjecture [16].)

It is interesting to note that one can show that the functions in (6) can be shown to satisfy the stated inequalities by establishing that the coefficients of their respective Maclaurin series, *expanded in powers of* ξ, satisfy the corresponding inequality relationships. In view of the preceding discussion, we now state the following proposition.

PROPOSITION 2.4 (see [9]). *Let G and A_p be as defined in Table* 1.1 *and let*

$$1+\sum_{n=1}^{\infty} b_n\xi^{2n} \equiv \left(\frac{1+(1-\xi^2)^{3/4}}{2}\right)^{2/3} \quad \textit{and} \tag{7}$$

$$1+\sum_{n=1}^{\infty} c_n\xi^{2n} \equiv (1-\xi^2/2)^{1/2}. \tag{8}$$

It follows that

$$b_n \le \frac{(1/2)_n(-1/2)_n}{n!\cdot n!} \le c_n \quad \textit{for all integers } n \ge 1.$$

Therefore, (6) *holds and is equivalent to $A_5(\lambda) \le G(\lambda) \le A_2(\lambda)$ for all $\lambda \in [0,1]$.*

Remark. If we apply the identity in (5) with $\lambda = (1-\sqrt{1-\xi^2})/(1+\sqrt{1-\xi^2})$, the definition of A_2, and simplify, we obtain $\Delta_2(\lambda) = 2[(1-\xi^2/2)^{1/2} - {}_2F_1(\frac{1}{2},-\frac{1}{2};1;\xi^2)]/(1+\sqrt{1-\xi^2})$. Proposition 2.4 implies that $(1-\xi^2/2)^{1/2} - {}_2F_1(\frac{1}{2},-\frac{1}{2};1;\xi^2)$ is a strictly increasing function of ξ. Therefore $\Delta_2(\lambda)$ is a strictly increasing function of ξ. Since $\xi = \frac{2\sqrt{\lambda}}{1+\lambda}$ is a strictly increasing function of λ on [0,1], it follows that $|\Delta_2|$ is a strictly increasing function of λ. A similar argument can be applied to $|\Delta_5|$.

Although some of the inequalities in the above propositions are straightforward, several proved to be surprisingly challenging to verify. In particular, the effort involving Almkvist's approximation A_9 precipitated the discovery of some deeper results involving the generalized hypergeometric function ${}_3F_2$, which are also of independent interest. In this direction, our main general results are as follows.

THEOREM 2.5. *Let $1 > a \ge b > -1$ and $1 > \varepsilon \ge \frac{(a+1)(b+2)}{a+b+4}$. Then $T_n \equiv {}_3F_2\left(-n,a,b;a+b+2,\varepsilon-n+1;1\right)$ satisfies*

$$ab(T_n - T_{n+1}) \ge 0 \quad \text{for all integers } n \ge 2.$$

COROLLARY 2.6. *Let $1 > a \ge b > -1$ and $1 > \varepsilon \ge \frac{(a+1)(b+2)}{a+b+4}$. Then $T_n \equiv {}_3F_2\left(-n,a,b;a+b+2,\varepsilon-n+1;1\right)$ satisfies*

$$abT_n \ge abT_{n+1} \ge ab\cdot {}_2F_1\left(a,b;a+b+2;1\right) \quad \text{for all integers } n \ge 2.$$

COROLLARY 2.7. *Let $1 > \varepsilon \ge \frac{1}{4}$. Then $T_n \equiv {}_3F_2\left(-n,-\frac{1}{2},-\frac{1}{2};1,\varepsilon-n+1;1\right)$ satisfies*

$$T_n \ge T_{n+1} \ge \frac{4}{\pi} \quad \text{for all integers } n \ge 2.$$

3. Verification of coefficient inequalities.

Proof of Proposition 2.2. *Part* I: *Almkvist's Approximation A_9.* Let $s \equiv (1-\lambda^2)^{1/2}$ and $\beta_n \equiv \beta_n^{(9)}$. It follows that

$$A_9(\lambda) = 2\left[\frac{(1+s)+(1-s)s}{(1+\sqrt{s})^2}\right] = \sum_{n=0}^{\infty}\beta_n\lambda^{2n},$$

which implies that

$$2(1+2s-s^2) = (1+2\sqrt{s}+s)\sum_{n=0}^{\infty}\beta_n\lambda^{2n}. \tag{9}$$

By replacing s by $(1-\lambda^2)^{1/2}$ and applying $(1-\lambda^2)^q = \sum_{n=0}^{\infty} \frac{(-q)_n}{n!}\lambda^{2n}$, we may change (9) to the form

$$2\lambda^2 + 4\sum_{n=0}^{\infty} \frac{(-1/2)_n}{n!}\lambda^{2n} = \sum_{n=0}^{\infty} \beta_n \lambda^{2n} + 2\sum_{n=0}^{\infty}\sum_{k=0}^{n} \frac{(-1/4)_{n-k}}{(n-k)!}\beta_k \lambda^{2n} + \sum_{n=0}^{\infty}\sum_{k=0}^{n} \frac{(-1/2)_{n-k}}{(n-k)!}\beta_k \lambda^{2n}.$$

Equating the coefficients of λ^{2n}, we obtain $\beta_0 = 1$, $\beta_1 = 1/4$, and

$$4\frac{(-1/2)_n}{n!} = \beta_n + 2\sum_{k=0}^{n} \frac{(-1/4)_{n-k}}{(n-k)!}\beta_k + \sum_{k=0}^{n} \frac{(-1/2)_{n-k}}{(n-k)!}\beta_k \quad \text{for } n \geq 2.$$

Solving for β_n, we have the recursive relationship

$$\beta_n = \frac{(-1/2)_n}{n!} - \frac{1}{2}\sum_{k=0}^{n-1} \frac{(-1/4)_{n-k}}{(n-k)!}\beta_k - \frac{1}{4}\sum_{k=0}^{n-1} \frac{(-1/2)_{n-k}}{(n-k)!}\beta_k \quad \text{for } n \geq 2. \tag{10}$$

We will use (10) and induction to show that

$$\beta_n \geq \alpha_n \quad \text{for all } n \geq 0. \tag{11}$$

First note that $\beta_n = \alpha_n$ for $n = 0, 1, 2$. Now let $n \geq 2$ and suppose that $\beta_k \geq \alpha_k$ for all $k = 0, \ldots, n-1$. Since the coefficients of β_k in (10) are all positive, it follows that

$$\beta_n \geq \frac{(-1/2)_n}{n!} - \frac{1}{2}\sum_{k=0}^{n-1} \frac{(-1/4)_{n-k}}{(n-k)!}\alpha_k - \frac{1}{4}\sum_{k=0}^{n-1} \frac{(-1/2)_{n-k}}{(n-k)!}\alpha_k.$$

Thus (11) will be established if we can verify that

$$\frac{(-1/2)_n}{n!} - \frac{1}{2}\sum_{k=0}^{n-1} \frac{(-1/4)_{n-k}}{(n-k)!}\alpha_k - \frac{1}{4}\sum_{k=0}^{n-1} \frac{(-1/2)_{n-k}}{(n-k)!}\alpha_k \geq \alpha_n \quad \text{for } n \geq 2. \tag{12}$$

Next we use the identities $(c)_{n-k} = \frac{(-1)^k (c)_n}{(1-c-n)_k}$ and $(1)_n = n!$ and add the corresponding nth term of each summation to both sides. Then (12) becomes

$$\frac{(-1/2)_n}{n!} - \frac{(-1/4)_n}{2 \cdot n!}\sum_{k=0}^{n} \frac{(-n)_k}{(5/4-n)_k}\alpha_k - \frac{(-1/2)_n}{4 \cdot n!}\sum_{k=0}^{n} \frac{(-n)_k}{(3/2-n)_k}\alpha_k \geq \frac{\alpha_n}{4}. \tag{13}$$

Now we apply $\alpha_k \equiv \left(\frac{(-1/2)_k}{k!}\right)^2$ and the definition of ${}_3F_2$, then divide both sides of (13) by $\frac{-(-1/2)_n}{4 \cdot n!}$, and simplify. Then inequality (13) becomes

$$P(n) \cdot {}_3F_2\left(-n, -\frac{1}{2}, -\frac{1}{2}; 1, \frac{5}{4} - n; 1\right) + {}_3F_2\left(-n, -\frac{1}{2}, -\frac{1}{2}; 1, \frac{3}{2} - n; 1\right) \geq Q(n), \tag{14}$$

where $P(n) \equiv 2\frac{(-1/4)_n}{(-1/2)_n}$ and $Q(n) \equiv 4 - \frac{(-1/2)_n}{n!}$. For $n \geq 2$, these can be shown to satisfy

$$P(n) \leq P(n+1) \quad \text{and} \tag{15}$$

$$Q(n) \geq Q(n+1). \tag{16}$$

We first note that inequality (14) can be confirmed directly for $n = 2, \ldots, 6$. An application of Corollary 2.7 (to be proved in the following section), with the respective values of $\varepsilon = 1/4$ and $\varepsilon = 1/2$, yields

$${}_3F_2\left(-n, -\frac{1}{2}, -\frac{1}{2}; 1, \frac{5}{4} - n; 1\right) \geq \frac{4}{\pi} \quad \text{and} \tag{17}$$

$${}_3F_2\left(-n, -\frac{1}{2}, -\frac{1}{2}; 1, \frac{3}{2} - n; 1\right) \geq \frac{4}{\pi} \tag{18}$$

for all $n \geq 2$. From inequalities (15)–(18) with $n \geq 6$, it follows that

$$P(n) \cdot {}_3F_2\left(-n, -\frac{1}{2}, -\frac{1}{2}; 1, \frac{5}{4} - n; 1\right) + {}_3F_2\left(-n, -\frac{1}{2}, -\frac{1}{2}; 1, \frac{3}{2} - n; 1\right)$$
$$\geq P(6)\frac{4}{\pi} + \frac{4}{\pi} \geq Q(6) \geq Q(n).$$

Therefore, inequality (14) holds for all $n \geq 2$ and hence $\beta_n^{(9)} \equiv \beta_n \geq \alpha_n$ for all $n \geq 0$. That is, Almkvist's approximation satisfies the property that all of the nonzero Maclaurin series coefficients of Δ_9 are positive. This concludes the proof of Part I of Proposition 2.2.

Proof of Proposition 2.2. *Part* II: *Jacobsen and Waadeland's Approximation* A_{12}. Now we seek to show that the approximation A_{12} satisfies the property that all of the nonzero Maclaurin series coefficients of Δ_{12} are negative. Let $a = 3$, $b = -112$, $c = 256$, and $D = \sqrt{b^2 - 4ac}$. It follows that

$$\frac{1}{au^2 + bu + c} = \frac{2a}{D}\left[\frac{1}{2au + b - D} - \frac{1}{2au + b + D}\right] = \sum_{n=0}^{\infty} d_n u^n \quad \text{for} \quad |u| < \left|\frac{D+b}{2a}\right|,$$

where

$$d_n \equiv \frac{2a}{D}\left[\frac{(-1)^n(2a)^n}{(b-D)^{n+1}} - \frac{(-1)^n(2a)^n}{(b+D)^{n+1}}\right] = \frac{1}{D}\left(\frac{2a}{D-b}\right)^{n+1}\left[\left(\frac{b-D}{b+D}\right)^{n+1} - 1\right].$$

It follows that $d_n > 0$ for all $n \geq 0$ and

$$\begin{aligned} A_{12}(\lambda) &\equiv \frac{256 - 48\lambda^2 - 21\lambda^4}{256 - 112\lambda^2 + 3\lambda^4} \\ &= -7 + \frac{2048 - 832\lambda^2}{256 - 112\lambda^2 + 3\lambda^4} \\ &= -7 + (2048 - 832\lambda^2)\sum_{n=0}^{\infty} d_n \lambda^{2n}. \end{aligned}$$

Now let $\beta_n \equiv \beta_n^{(12)}$. Then the nonzero Maclaurin series coefficients for A_{12} are given by $\beta_0 = 1$ and

$$\beta_n = 2048 d_n - 832 d_{n-1} \quad \text{for all } n \geq 1.$$

Since $(x^{n+1}-1)/(x-1) > x$ for $x \equiv (b-D)/(b+D) > 1$, it follows easily that $(2048d_n)/(832d_{n-1}) > 1$ for all $n \geq 1$. Thus

$$\beta_n > 0 \quad \text{for all } n \geq 0. \tag{19}$$

Direct calculation reveals that $\beta_n = \alpha_n$ for $n = 0, \ldots, 4$. Also note that

$$(256 - 112\lambda^2 + 3\lambda^4) \sum_{n=0}^{\infty} \beta_n \lambda^{2n} = 256 - 48\lambda^2 - 21\lambda^4.$$

Hence

$$\sum_{n=3}^{\infty} (256\beta_n - 112\beta_{n-1} + 3\beta_{n-2})\lambda^{2n} = 0. \tag{20}$$

Thus the coefficients of λ^{2n} in (20) are zero for all $n \geq 3$. Solving for β_n and using (19), we have

$$\beta_n = (112\beta_{n-1} - 3\beta_{n-2})/256 < \frac{112}{256}\beta_{n-1} \quad \text{for all } n \geq 3.$$

Now suppose that $\beta_n \leq \alpha_n$ for some integer $n \geq 4$, where $\alpha_n \equiv (\frac{(-1/2)_n}{n!})^2$. Then

$$\beta_{n+1} < \frac{112}{256}\beta_n \leq \frac{112}{256}\alpha_n = \frac{112}{256}\frac{\alpha_n}{\alpha_{n+1}}\alpha_{n+1} = \frac{112}{256}\left(\frac{n+1}{n-\frac{1}{2}}\right)^2 \alpha_{n+1} \leq \alpha_{n+1}.$$

Thus $\beta_n^{(12)} \equiv \beta_n \leq \alpha_n$ for all integers $n \geq 0$. This concludes the proof of Part II of Proposition 2.2. □

Before proving Proposition 2.3, we first observe that the nine approximations involved have the following respective Maclaurin series representations (recursive relationships satisfied by $\beta_n^{(13)}$ and $\beta_n^{(3)}$ are developed in the appendix):

$$A_1(\lambda) \equiv (1-\lambda^2)^{1/2} = 1 + \sum_{n=1}^{\infty} \frac{(-1/2)_n}{n!}\lambda^{2n}, \tag{21}$$

$$A_6(\lambda) \equiv \left(1 + \frac{\lambda^2}{8}\right)^2 = 1 + \frac{\lambda^2}{4} + \frac{\lambda^4}{64}, \tag{22}$$

$$A_7(\lambda) \equiv 1 + \frac{\lambda^2/4}{1-\lambda^2/16} = 1 + \frac{\lambda^2}{4} + \sum_{n=2}^{\infty} \frac{1}{2^{4n-2}}\lambda^{2n}, \tag{23}$$

$$A_8(\lambda) \equiv 3 - \sqrt{4-\lambda^2} = 1 + \frac{\lambda^2}{4} - \sum_{n=2}^{\infty} \frac{(-1/2)_n}{n!2^{2n-1}}\lambda^{2n}, \tag{24}$$

$$A_{10}(\lambda) \equiv \frac{64 - 3\lambda^4}{64 - 16\lambda^2} = 1 + \frac{\lambda^2}{4} + \sum_{n=2}^{\infty} \frac{1}{2^{2n+2}}\lambda^{2n}, \tag{25}$$

$$A_{11}(\lambda) \equiv \frac{3}{2} + \frac{\lambda^2}{8} - \frac{1}{2}\left(1 - \frac{\lambda^2}{2}\right)^{1/2} = 1 + \frac{\lambda^2}{4} - \sum_{n=2}^{\infty} \frac{(-1/2)_n}{n!2^{n+1}}\lambda^{2n}, \tag{26}$$

$$A_{13}(\lambda) \equiv 1 + \frac{3\lambda^2}{10 + \sqrt{4 - 3\lambda^2}} = 1 + \frac{\lambda^2}{4} + \sum_{n=2}^{\infty} \beta_n^{(13)} \lambda^{2n}, \tag{27}$$

$$A_4(\lambda) \equiv \frac{3}{2} - \frac{1}{2}(1 - \lambda^2)^{1/2} = 1 + \frac{\lambda^2}{4} - \frac{1}{2} \sum_{n=2}^{\infty} \frac{(-1/2)_n}{n!} \lambda^{2n}, \tag{28}$$

$$A_3(\lambda) \equiv \frac{2}{1 + \sqrt{1 - \lambda^2}} = 1 + \frac{\lambda^2}{4} + \sum_{n=2}^{\infty} \beta_n^{(3)} \lambda^{2n}. \tag{29}$$

Proof of Proposition 2.3. We seek to establish the following inequalities regarding the specified Maclaurin series coefficients:

$$\beta_n^{(1)} \le \beta_n^{(6)} \le \beta_n^{(7)} \le \beta_n^{(8)} \le \beta_n^{(10)} \le \beta_n^{(11)} \le \beta_n^{(13)} \le \alpha_n \le \beta_n^{(4)} \le \beta_n^{(3)} \tag{30}$$

for all $n \ge 0$. Referring to (21)–(29), we note that the inequalities in (30) are trivial for $n = 0$ and $n = 1$. Thus we must verify (30) for all $n \ge 2$. The first two inequalities are immediate while the next three inequalities follow directly by induction. We now proceed to prove the remaining inequalities in (30).

• Claim I. $\beta_n^{(11)} \le \beta_n^{(13)} \le \alpha_n$ *for all* $n \ge 2$.

Let $\beta_n \equiv \beta_n^{(13)}$ and $\gamma_n \equiv \beta_n^{(11)}$, where $\beta_n^{(11)} \equiv \frac{-(-1/2)_n}{n! 2^{n+1}}$ for $n \ge 2$ (see (26)) and recall that $\alpha_n \equiv (\frac{(-1/2)_n}{n!})^2$. The nonzero Maclaurin series coefficients of Ramanujan's second estimate A_{13} can be shown to satisfy (see the appendix) $\beta_0 = 1$, $\beta_1 = 1/4$, $\beta_2 = 1/64$, and

$$\beta_n = \phi_{n-1} - 2^{-5}\beta_{n-1} \quad \text{for all } n \ge 3, \quad \text{where} \quad \phi_n \equiv -\frac{(-1/2)_n (3/4)^n}{16 \cdot n!}. \tag{31}$$

Applying (31) twice, we have

$$\beta_n = \phi_{n-1} - 2^{-5}\phi_{n-2} + 2^{-10}\beta_{n-2} \quad \text{for all } n \ge 4. \tag{32}$$

Direct calculation reveals that Claim I holds for $n = 2, 3, 4$. That is, $\gamma_n \le \beta_n \le \alpha_n$ for $n = 2, 3, 4$. Now let $n \ge 5$ and suppose that

$$\gamma_k \le \beta_k \le \alpha_k \quad \text{for all } k = 2, \ldots, n-1. \tag{33}$$

Then (32) and (33) together imply that

$$\begin{aligned} &\phi_{n-1} - 2^{-5}\phi_{n-2} + 2^{-10}\gamma_{n-2} \\ &\le \overbrace{\phi_{n-1} - 2^{-5}\phi_{n-2} + 2^{-10}\beta_{n-2}}^{\beta_n} \le \phi_{n-1} - 2^{-5}\phi_{n-2} + 2^{-10}\alpha_{n-2}. \end{aligned} \tag{34}$$

It can be shown (see the appendix) that

$$\gamma_n \le \phi_{n-1} - 2^{-5}\phi_{n-2} + 2^{-10}\gamma_{n-2} \quad \text{and} \tag{35}$$

$$\alpha_n \ge \phi_{n-1} - 2^{-5}\phi_{n-2} + 2^{-10}\alpha_{n-2} \tag{36}$$

for all $n \ge 5$. Therefore, using inequalities (34)–(36) and induction, we have $\gamma_n \le \beta_n \le \alpha_n$ for all $n \ge 2$. This completes the proof of Claim I.

• Claim II. $\alpha_n \le \beta_n^{(4)} \le \beta_n^{(3)}$ *for all* $n \ge 2$.

If we now apply (28), the first inequality in Claim II becomes

$$\alpha_n \equiv \left(\frac{(-1/2)_n}{n!}\right)^2 \leq \frac{-(-1/2)_n}{2 \cdot n!} \equiv \beta_n^{(4)} \quad \text{for all } n \geq 2.$$

This is equivalent to

$$\frac{-2(-1/2)_n}{n!} \leq 1 \quad \text{for all } n \geq 2$$

which follows by induction. The second inequality in Claim II involves the Maclaurin series coefficients of Sipos and Ekwall's approximation A_3 which can be shown to satisfy the following recursive relationship (see the appendix): $\beta_0^{(3)} = 1$, $\beta_1^{(3)} = 1/4$, $\beta_2^{(3)} = 1/8$, and

$$\beta_n^{(3)} = -\frac{1}{2}\sum_{k=0}^{n-1} \frac{(-1/2)_{n-k}}{(n-k)!}\beta_k^{(3)} \quad \text{for all } n \geq 2. \tag{37}$$

Note that

$$-\frac{1}{2}\sum_{k=0}^{n-1} \frac{(-1/2)_{n-k}}{(n-k)!}\beta_k^{(3)} = \frac{-(-1/2)_n}{2 \cdot n!} - \frac{1}{2}\sum_{k=1}^{n-1} \frac{(-1/2)_{n-k}}{(n-k)!}\beta_k^{(3)} \tag{38}$$

for all $n \geq 2$, and

$$\frac{-(-1/2)_{n-k}}{2 \cdot (n-k)!}\beta_k^{(3)} > 0 \quad \text{for } k = 1, \ldots, n-1. \tag{39}$$

Therefore, (37)–(39) together yield

$$\beta_n^{(4)} \equiv \frac{-(-1/2)_n}{2 \cdot n!} \leq \beta_n^{(3)} \quad \text{for all } n \geq 2.$$

This concludes the proof of Claim II and Proposition 2.3. □

Remarks on the Proof of Proposition 2.4. From (8), we have that $c_n \equiv \dfrac{(1/2)^n(-1/2)_n}{n!}$ for all $n \geq 1$. By induction, it can be shown that

$$\frac{(1/2)_n(-1/2)_n}{n! \cdot n!} \leq c_n \quad \text{for all } n \geq 1.$$

In an earlier paper (see [9]), the authors use the logarithmic derivative and Cauchy products to obtain the recursive relationship for b_n (with b_n as defined in (7)) given by

$$b_{n+1} = \frac{1}{2(n+1)}\left[\left(\frac{5}{4}n - \frac{1}{2}\right)b_n - \sum_{k=0}^{n-2}(k+1)b_{k+1}\frac{\left(-\frac{1}{4}\right)_{n-k}}{(n-k)!}\right]. \tag{40}$$

Theorem 2.1, together with (40), was then used (see [9]) to establish that

$$b_n \leq \frac{(1/2)_n(-1/2)_n}{n! \cdot n!} \quad \text{for all } n \geq 1. \qquad \square$$

4. Proofs of general results involving ${}_3F_2$. We will make use of the following classical identities which we include for the reader's convenience ($F \equiv {}_3F_2$).

IDENTITY 1 {see [13, p. 440, eq. (33)]}.

$$F(\rho,a,b;c,\sigma;1) - F(\rho+1,a,b;c,\sigma+1;1) = \frac{-ab(\sigma-\rho)}{c\sigma(\sigma+1)} \cdot F(\rho+1,a+1,b+1;c+1,\sigma+2;1).$$

IDENTITY 2 {see [11, p. 59, eq. (3.1.1)]}.

$$F(-n,a,b;c,d;1) = \frac{(d-b)_n}{(d)_n} \cdot F(-n,c-a,b;c,1+b-d-n;1).$$

IDENTITY 3 {see [13, p. 440, eq. (26)]}.

$$\sigma \cdot F(\rho,a,b;c,\sigma;1) = \rho \cdot F(\rho+1,a,b;c,\sigma+1;1) + (\sigma-\rho) \cdot F(\rho,a,b;c,\sigma+1;1).$$

IDENTITY 4 {see [14, p. 82, eq. (14)]}.

$$(a_1-a_2)\cdot F(a_1,a_2,a_3;b_1,b_2;z) = a_1 \cdot F(a_1+1,a_2,a_3;b_1,b_2;z) - a_2 \cdot F(a_1,a_2+1,a_3;b_1,b_2;z).$$

IDENTITY 5 {see [13, p. 440, eq. (30)]}.

$$F(\sigma,a,b;c,d;1) - F(\sigma+1,a,b;c,d;1) = \frac{-ab}{cd} \cdot F(\sigma+1,a+1,b+1;c+1,d+1;1).$$

Proof of Theorem 2.5. Define $T_n \equiv F(-n,a,b;a+b+2,\varepsilon-n+1;1)$, where $F \equiv {}_3F_2$. Let $1 > a \geq b > -1$ and $1 > \varepsilon \geq \frac{(a+1)(b+2)}{a+b+4}$. For $n \geq 2$, it follows that

$$\begin{aligned}
T_{n+1} - T_n &= F(-n-1,a,b;a+b+2,\varepsilon-n;1) - F(-n,a,b;a+b+2,\varepsilon-n+1;1) \\
&= \frac{-ab(\varepsilon+1)}{(\varepsilon-n)(\varepsilon-n+1)(a+b+2)} F(-n,a+1,b+1;a+b+3,\varepsilon-n+2;1) \\
&\quad \{\text{using Identity 1 with } \rho = -n-1,\ \sigma = \varepsilon-n\} \\
&= \frac{-ab(\varepsilon+1)}{(n-\varepsilon)(n-\varepsilon-1)(a+b+2)} \frac{(\varepsilon-n-b+1)_n}{(\varepsilon-n+2)_n} \\
&\quad \times F(-n,b+2,b+1;a+b+3,b-\varepsilon;1) \\
&\quad \{\text{using Identity 2}\} \\
&= \frac{-ab(\varepsilon+1)(b-\varepsilon)_n}{(n-\varepsilon)(n-\varepsilon-1)(a+b+2)(-1-\varepsilon)_n(b-\varepsilon)} \\
&\quad \times [(b+1)F(-n,b+2,b+2;a+b+3,b+1-\varepsilon;1) \\
&\quad + (-\varepsilon-1)F(-n,b+2,b+1;a+b+3,b+1-\varepsilon;1)], \qquad (41)
\end{aligned}$$

where (41) follows from Identity 3 (with $\rho = b+1$, $\sigma = b-\varepsilon$) and the identity $(1-\alpha-n)_n = (-1)^n(\alpha)_n$.
Identity 4 (with $a_1 = -n$ and $a_2 = b+1$) implies that

$$\begin{aligned}
&F(-n,b+2,b+2;a+b+3,b+1-\varepsilon;1) \\
&\quad = \frac{1}{b+1}[(n+b+1)F(-n,b+2,b+1;a+b+3,b+1-\varepsilon;1) \\
&\quad + (-n)F(-n+1,b+2,b+1;a+b+3,b+1-\varepsilon;1)]. \qquad (42)
\end{aligned}$$

Now let $G_n = F(-n, b+2, b+1; a+b+3, b+1-\varepsilon; 1)$ and use (41) and (42). Then we have that

$$\begin{aligned} T_{n+1} - T_n &= \frac{-ab(\varepsilon+1)(b-\varepsilon)_n}{(n-\varepsilon)(n-\varepsilon-1)(a+b+2)(-1-\varepsilon)_n(b-\varepsilon)} \\ &\quad \times[(n+b+1)G_n - nG_{n-1} + (-\varepsilon-1)G_n] \\ &= \frac{-ab(\varepsilon+1)(b-\varepsilon)_n}{(n-\varepsilon)(n-\varepsilon-1)(a+b+2)(-1-\varepsilon)_n(b-\varepsilon)} \\ &\quad \times[n(G_n - G_{n-1}) + (b-\varepsilon)G_n]. \end{aligned} \tag{43}$$

Applications of Identity 5 (with $\sigma = -n$) followed by Identity 2 yield

$$\begin{aligned} G_n - G_{n-1} &= F(-n, b+2, b+1; a+b+3, b+1-\varepsilon; 1) \\ &\quad - F(-n+1, b+2, b+1; a+b+3, b+1-\varepsilon; 1) \\ &= \frac{-(b+2)(b+1)}{(a+b+3)(b+1-\varepsilon)} F(-n+1, b+3, b+2; a+b+4, b+2-\varepsilon; 1) \\ &= \frac{-(b+2)(b+1)}{(a+b+3)(b+1-\varepsilon)} \cdot \frac{(-\varepsilon)_{n-1}}{(b+2-\varepsilon)_{n-1}} \\ &\quad \times F(-n+1, a+1, b+2; a+b+4, \varepsilon-n+2; 1). \end{aligned} \tag{44}$$

Identity 2 also implies that

$$G_n = \frac{(-\varepsilon)_n}{(b+1-\varepsilon)_n} F(-n, a+1, b+1; a+b+3, \varepsilon-n+1; 1). \tag{45}$$

Combining (43)–(45), we have

$$\begin{aligned} T_{n+1} - T_n &= \frac{-ab(\varepsilon+1)(b-\varepsilon)_n}{(n-\varepsilon)(n-\varepsilon-1)(a+b+2)(-1-\varepsilon)_n(b-\varepsilon)} \\ &\times \left[\frac{-n(b+2)(b+1)(-\varepsilon)_{n-1}}{(a+b+3)(b+1-\varepsilon)(b+2-\varepsilon)_{n-1}} F(-n+1, a+1, b+2; a+b+4, \varepsilon-n+2; 1) \right. \\ &\quad \left. + (b-\varepsilon) \frac{(-\varepsilon)_n}{(b+1-\varepsilon)_n} F(-n, a+1, b+1; a+b+3, \varepsilon-n+1; 1) \right]. \end{aligned} \tag{46}$$

Now make use of $\frac{(b-\varepsilon)_n}{(b+1-\varepsilon)_n} = \frac{(b-\varepsilon)}{(n+b-\varepsilon)}$, $\frac{(-\varepsilon)_n}{(-1-\varepsilon)_n} = \frac{(-1-\varepsilon+n)}{(-1-\varepsilon)}$, $\frac{(-\varepsilon)_{n-1}}{(-1-\varepsilon)_n} = \frac{1}{(-1-\varepsilon)}$, and multiply both sides by $-ab$. Then (46) becomes

$$\begin{aligned} ab(T_n - T_{n+1}) &= \frac{(ab)^2(\varepsilon+1)}{(n-\varepsilon)(n-\varepsilon-1)(a+b+2)(b-\varepsilon)} \\ &\times \left[\frac{-n(b+2)(b+1)(b-\varepsilon)}{(a+b+3)(-1-\varepsilon)(n+b-\varepsilon)} F(-n+1, a+1, b+2; a+b+4, \varepsilon-n+2; 1) \right. \\ &\quad \left. + (b-\varepsilon) \frac{(n-1-\varepsilon)(b-\varepsilon)}{(-1-\varepsilon)(n+b-\varepsilon)} F(-n, a+1, b+1; a+b+3, \varepsilon-n+1; 1) \right] \\ &= \frac{(ab)^2}{(n-\varepsilon)(n-\varepsilon-1)(a+b+2)(n+b-\varepsilon)} \\ &\quad \times \left[\frac{n(b+2)(b+1)}{(a+b+3)} F(-(n-1), a+1, b+2; a+b+4, \varepsilon-(n-1)+1; 1) \right. \\ &\quad \left. + (\varepsilon-b)(n-\varepsilon-1) F(-n, a+1, b+1; a+b+3, \varepsilon-n+1; 1) \right], \end{aligned} \tag{47}$$

where $n+b-\varepsilon > n-\varepsilon-1 > n-2 \geq 0$, $n-\varepsilon > 0$, and $\varepsilon - b > \varepsilon - \frac{(a+1)(b+2)}{a+b+4} \geq 0$. Since $1 > \varepsilon \geq \frac{(a+1)(b+2)}{a+b+4} > \frac{(a+1)(b+1)}{a+b+3}$, Theorem 2.1 implies that

$$F\left(-(n-1), a+1, b+2; a+b+4, \varepsilon-(n-1)+1; 1\right) \geq 0 \quad \text{and}$$
$$F\left(-n, a+1, b+1; a+b+3, \varepsilon-n+1; 1\right) \geq 0.$$

Therefore, (47) is the product and sum of nonnegative quantities and thus

$$ab(T_n - T_{n+1}) \geq 0 \quad \text{for all integers } n \geq 2. \qquad \square$$

In order to prove Corollary 2.6, we will make use of the following two lemmas.

LEMMA 4.1. *Let n be a positive integer and $0 < \varepsilon < 1$. Then*

$$\frac{(-n)_k}{(\varepsilon-n+1)_k} \geq 1 \quad \textit{for all } k = 0, \ldots, n-1.$$

Proof of Lemma 4.1. Note that the desired inequality holds at $k = 0$. Now let $n \geq 2$ and suppose that

$$\frac{(-n)_k}{(\varepsilon-n+1)_k} \geq 1 \quad \text{for some } k \text{ with } 0 \leq k \leq n-2.$$

Then

$$\frac{(-n)_{k+1}}{(\varepsilon-n+1)_{k+1}} = \frac{(-n)_k(-n+k)}{(\varepsilon-n+1)_k(\varepsilon-n+1+k)} \geq \frac{(-n)_k}{(\varepsilon-n+1)_k} \geq 1. \qquad \square$$

LEMMA 4.2. *Define $\psi_n(a,b,c,\varepsilon) \equiv \frac{(a)_n(b)_n(-n)_n}{n!(c)_n(\varepsilon-n+1)_n}$. Let (a,b,c,ε) be in the domain of ψ_n for all $n \geq 2$ with $\varepsilon < c-a-b$. Then*

$$\lim_{n\to\infty} \psi_n(a,b,c,\varepsilon) = 0.$$

Proof of Lemma 4.2. Since $(1-c-n)_n = (-1)^n(c)_n$, it follows that

$$\psi_n = \frac{(a)_n(b)_n(1)_n}{n!(c)_n(-\varepsilon)_n} = \frac{\Gamma(a+n)\Gamma(b+n)\Gamma(c)\Gamma(-\varepsilon)}{\Gamma(a)\Gamma(b)\Gamma(c+n)\Gamma(-\varepsilon+n)} n^{c-a-b-\varepsilon} n^{a+b+\varepsilon-c}.$$

It is known that (see [1, p. 257, eq. (6.1.46)])

$$\lim_{n\to\infty} \frac{\Gamma(r+n)}{\Gamma(s+n)} n^{s-r} = 1.$$

If $a+b+\varepsilon-c < 0$, then

$$\lim_{n\to\infty} \psi_n = \lim_{n\to\infty} \frac{\Gamma(a+n)\Gamma(b+n)\Gamma(c)\Gamma(-\varepsilon)}{\Gamma(a)\Gamma(b)\Gamma(c+n)\Gamma(-\varepsilon+n)} n^{c-a-b-\varepsilon} \cdot \lim_{n\to\infty} n^{a+b+\varepsilon-c} = 0. \qquad \square$$

Proof of Corollary 2.6. Let $1 > a \geq b > -1$ and $1 > \varepsilon \geq \frac{(a+1)(b+2)}{a+b+4}$ and define

$$T_n \equiv {}_3F_2\left(-n, a, b; a+b+2, \varepsilon-n+1; 1\right).$$

Theorem 2.5 implies that the sequence $\{abT_n\}_{n=2}^{\infty}$ is a monotone (nonincreasing) sequence. Now define

$$S_n \equiv 1 + \frac{(a)_n(b)_n(-n)_n}{n!(a+b+2)_n(\varepsilon-n+1)_n} + \sum_{k=1}^{n-1} \frac{(a)_k(b)_k}{k!(a+b+2)_k}.$$

Using the definition of ${}_3F_2$, Lemma 4.1, and the fact that $\dfrac{ab(a)_k(b)_k}{k!(a+b+2)_k} \geq 0$ for $k = 1, \ldots, n-1$, we obtain

$$abT_n = ab + \frac{ab(a)_n(b)_n(-n)_n}{n!(a+b+2)_n(\varepsilon-n+1)_n} + \sum_{k=1}^{n-1} \frac{ab(a)_k(b)_k(-n)_k}{k!(a+b+2)_k(\varepsilon-n+1)_k} \geq abS_n$$

for all $n \geq 2$.
Applying Lemma 4.2 with $c = a+b+2$, we have

$$\lim_{n\to\infty} S_n = {}_2F_1\left(a, b; a+b+2; 1\right).$$

Since $abT_n \geq abS_n$ for all $n \geq 2$, it follows that $\{abT_n\}_{n=2}^{\infty}$ is a bounded monotone sequence. Thus

$$abT_n \geq \lim_{n\to\infty} abT_n \geq \lim_{n\to\infty} abS_n = ab \cdot {}_2F_1\left(a, b; a+b+2; 1\right) \quad \text{for all } n \geq 2. \qquad \square$$

Proof of Corollary 2.7. Choose $a = b = -1/2$ and $1 > \varepsilon \geq 1/4$ and define

$$T_n \equiv {}_3F_2\left(-n, -\frac{1}{2}, -\frac{1}{2}; 1, \varepsilon-n+1; 1\right).$$

It is known that (see [14, p. 49])

$${}_2F_1\left(-\frac{1}{2}, -\frac{1}{2}; 1; 1\right) = \frac{4}{\pi}.$$

Corollary 2.6 implies that

$$T_n \geq T_{n+1} \geq \frac{4}{\pi} \quad \text{for all } n \geq 2. \qquad \square$$

5. Appendix.

5.1. Recursive relationship for Maclaurin series coefficients of Ramanujan's second estimate A_{13}. Writing $\beta_n \equiv \beta_n^{(13)}$, we have

$$3\lambda^2(10 - \sqrt{4-3\lambda^2}) = (A_{13}(\lambda) - 1)(10^2 - (\sqrt{4-3\lambda^2})^2) = (96 + 3\lambda^2)\sum_{n=1}^{\infty} \beta_n \lambda^{2n}$$

which implies that

$$10 - 2\left(1 - \frac{3}{4}\lambda^2\right)^{1/2} = (32 + \lambda^2)\sum_{n=1}^{\infty} \beta_n \lambda^{2n-2}.$$

Applying $(1-x)^q = \sum_{n=0}^{\infty} \frac{(-q)_n}{n!} x^n$ and simplifying yields

$$8 - 2\sum_{n=1}^{\infty} \frac{(-1/2)_n (3/4)^n}{n!} \lambda^{2n} = 32\beta_1 + \sum_{n=1}^{\infty} (32\beta_{n+1} + \beta n)\lambda^{2n}.$$

Thus $\beta_0 = 1$, $\beta_1 = 1/4$, $\beta_2 = 1/64$, and

$$\beta_{n+1} = \frac{-(-1/2)_n (3/4)^n}{16 \cdot n!} - \frac{\beta_n}{32} \quad \text{for all } n \geq 1.$$

Letting $\phi_n \equiv -\dfrac{(-1/2)_n (3/4)^n}{16 \cdot n!}$, we obtain

$$\beta_{n+1} = \phi_n - 2^{-5}\beta_n \quad \text{for all } n \geq 2. \qquad \square$$

5.2. Recursive relationship for Maclaurin series coefficients of Sipos and Ekwall's estimate A_3. Writing $\beta_n \equiv \beta_n^{(3)}$ and using the Cauchy product, we have

$$2 = A_3(\lambda)(1 + \sqrt{1-\lambda^2}) = \sum_{n=0}^{\infty} \beta_n \lambda^{2n} + \sum_{n=0}^{\infty} \sum_{k=0}^{n} \frac{(-1/2)_{n-k}}{(n-k)!} \beta_k \lambda^{2n}.$$

Thus $\beta_0^{(3)} = 1,\ \beta_1^{(3)} = 1/4,\ \beta_2^{(3)} = 1/8,$ and

$$\beta_n^{(3)} = \frac{-1}{2} \sum_{k=0}^{n-1} \frac{(-1/2)_{n-k}}{(n-k)!} \beta_k^{(3)} \quad \text{for all } n \geq 2. \qquad \square$$

5.3. Establishing inequality (35). Let $\phi_n \equiv \frac{-(-1/2)_n (3/4)^n}{16 \cdot n!}$, $\gamma_n \equiv \beta_n^{(11)} = \frac{-(-1/2)_n}{n! 2^{n+1}}$, and $n \geq 4$. Inequality (35) claims that $\gamma_n \leq \phi_{n-1} - 2^{-5}\phi_{n-2} + 2^{-10}\gamma_{n-2}$.

Direct calculation reveals that the desired inequality holds for $n = 4, \ldots, 7$. Now suppose that $n \geq 7$. Since $\gamma_{n-2} > 0$, we have

$$\begin{aligned}
&\gamma_n - \phi_{n-1} + 2^{-5}\phi_{n-2} - 2^{-10}\gamma_{n-2} < \gamma_n - \phi_{n-1} + 2^{-5}\phi_{n-2} \\
&\quad = \frac{-(-1/2)_n}{n! 2^{n+1}} + \frac{(-1/2)_{n-1}(3/4)^{n-1}}{16 \cdot (n-1)!} - 2^{-9} \frac{(-1/2)_{n-2}(3/4)^{n-2}}{(n-2)!} \\
&\quad = -2^{-9} \frac{(-1/2)_{n-2}(3/4)^{n-2}}{(n-2)!} \cdot \left[\frac{2^9 (4/3)^{n-2}(n-3/2)(n-5/2)}{n(n-1)2^{n+1}} - \frac{2^5 (3/4)(n-5/2)}{(n-1)} + 1 \right] \\
&\quad = -2^{-9} \frac{(-1/2)_{n-2}(3/4)^{n-2}}{(n-2)!} \cdot \left[\frac{2^{n+4}(n-3/2)(n-5/2)}{3^{n-2} n(n-1)} - \frac{24(n-5/2)}{(n-1)} + 1 \right].
\end{aligned}$$

Since $\frac{(n-5/2)}{(n-1)} \geq \frac{1}{2}$, it follows that

$$\frac{2^{n+4}(n-3/2)(n-5/2)}{3^{n-2} n(n-1)} - \frac{24(n-5/2)}{(n-1)} + 1 \leq \frac{2^{n+4}(n-3/2)(n-5/2)}{3^{n-2} n(n-1)} - 11.$$

Thus

$$\begin{aligned}
&\gamma_n - \phi_{n-1} + 2^{-5}\phi_{n-2} - 2^{-10}\gamma_{n-2} \\
&\quad < -2^{-5} \cdot 9 \frac{(-1/2)_{n-2}(3/4)^{n-2}}{(n-2)!} \cdot \left[\frac{2^n (n-3/2)(n-5/2)}{3^n n(n-1)} - \frac{11 \cdot 2^{-4}}{3^2} \right]
\end{aligned}$$

$$\leq \frac{-9(-1/2)_{n-2}(3/4)^{n-2}}{32(n-2)!}\cdot\left[\left(\frac{2}{3}\right)^n-\frac{11}{144}\right]$$
$$\leq \frac{-9(-1/2)_{n-2}(3/4)^{n-2}}{32(n-2)!}\cdot\left[\left(\frac{2}{3}\right)^7-\frac{11}{144}\right]<0.$$

Hence the claim in (35) is established. □

5.4. Establishing inequality (36). Let $\phi_n \equiv -\frac{(-1/2)_n(3/4)^n}{16\cdot n!}$, $\alpha_n \equiv (\frac{(-1/2)_n}{n!})^2$, and $n \geq 4$.

Inequality (36) claims that $\phi_n - 2^{-5}\phi_{n-1} + 2^{-10}\alpha_{n-1} \leq \alpha_{n+1}$. Note that

$$\begin{aligned}
&\phi_{\ n} - 2^{-5}\phi_{n-1} + 2^{-10}\alpha_{n-1} - \alpha_{n+1}\\
&= -\frac{(-1/2)_n(3/4)^n}{16\cdot n!} + 2^{-5}\frac{(-1/2)_{n-1}(3/4)^{n-1}}{16\cdot(n-1)!} + 2^{-10}\left(\frac{(-1/2)_{n-1}}{(n-1)!}\right)^2 - \left(\frac{(-1/2)_{n+1}}{(n+1)!}\right)^2\\
&= \frac{(-1/2)_{n-1}}{(n-1)!}\left\{\frac{-(n-3/2)3^n}{n2^{2n+4}} + \frac{3^{n-1}}{2^{2n+7}} + \frac{(-1/2)_{n-1}}{2^{10}(n-1)!} - \frac{(n-3/2)(n-1/2)(-1/2)_{n+1}}{n(n+1)\cdot(n+1)!}\right\}\\
&= \frac{(-1/2)_{n-1}}{(n-1)!}\left\{\frac{3^{n-1}}{2^{2n+4}}\left[\frac{3(3/2-n)}{n}+\frac{1}{2^3}\right] + \frac{(-1/2)_{n-1}}{(n-1)!}\left[\frac{1}{2^{10}} - \frac{(n-3/2)^2(n-1/2)^2}{n^2(n+1)^2}\right]\right\}\\
&= \frac{(-1/2)_{n-1}}{(n-1)!}\left\{\frac{U(n)}{V(n)}+1\right\}V(n),
\end{aligned}$$

where

$$U(n) \equiv \frac{3^{n-1}}{2^{2n+4}}\left[\frac{3(3/2-n)}{n}+\frac{1}{2^3}\right]$$

and

$$V(n) \equiv \frac{(-1/2)_{n-1}}{(n-1)!}\left[\frac{1}{2^{10}} - \frac{(n-3/2)^2(n-1/2)^2}{n^2(n+1)^2}\right].$$

It follows that $V(n) > 0$. Now let $W(n) \equiv U(n)/V(n)$. Since $(-1/2)_{n-1} < 0$, we will be finished if we can show that $W(n)+1 > 0$ for all $n \geq 4$. Direct calculation again yields $W(4)+1 > 0$. For $n \geq 4$, it is easy to check that

$$\begin{aligned}
W(n+1) - W(n) &= \frac{\frac{3^n}{2^{2n+6}}\left[\frac{3(1/2-n)}{n+1}+\frac{1}{2^3}\right]}{\frac{(-1/2)_n}{n!}\left[\frac{1}{2^{10}} - \frac{(n-1/2)^2(n+1/2)^2}{(n+1)^2(n+2)^2}\right]}\\
&\quad - \frac{\frac{3^{n-1}}{2^{2n+4}}\left[\frac{3(3/2-n)}{n}+\frac{1}{2^3}\right]}{\frac{(-1/2)_{n-1}}{(n-1)!}\left[\frac{1}{2^{10}} - \frac{(n-3/2)^2(n-1/2)^2}{n^2(n+1)^2}\right]}\\
&= \left\{\frac{\frac{3^{n-1}}{2^{2n+4}}}{\frac{(-1/2)_{n-1}}{(n-1)!}}\right\}\left\{\frac{3n}{4(n-3/2)}Z(n+1) - Z(n)\right\}, \qquad (48)
\end{aligned}$$

where $Z(n) \equiv [\frac{3(3/2-n)}{n}+\frac{1}{2^3}]/[\frac{1}{2^{10}} - \frac{(n-3/2)^2(n-1/2)^2}{n^2(n+1)^2}]$. Direct calculation reveals that the expression in (48) is nonnegative for $n = 4$ and $n = 5$. For $n \geq 6$, it can be shown by a straightforward calculation that $0 < Z(n+1) \leq Z(n)$. Hence $\frac{3n}{4(n-3/2)}Z(n+1) - Z(n) \leq Z(n+1) - Z(n) \leq 0$ for all $n \geq 6$. Thus $W(n+1) - W(n) \geq 0$ for all $n \geq 4$ since $(-1/2)_{n-1} < 0$. Therefore, $W(n)+1 \geq W(4)+1 > 0$ for all $n \geq 4$. This establishes the claim in (36). □

Acknowledgments. The authors wish to acknowledge the referees for their helpful suggestions regarding the revision of this paper. The authors also wish to thank H. Alzer for useful correspondence.

REFERENCES

[1] M. Abramowitz and I.A. Stegun, *Handbook of Mathematical Functions*, Dover, New York, 1972.
[2] G. Almkvist and B. Berndt, *Gauss, Landen, Ramanujan, the arithmetic-geometric mean, ellipses, π, and the Ladies Diary*, Amer. Math. Monthly, 95 (1988), pp. 585–608.
[3] H. Alzer, *private communication*, 1998.
[4] G.D. Anderson, R.W. Barnard, K.C. Richards, M.K. Vamanamurthy, and M. Vuorinen, *Inequalities for zero-balanced hypergeometric functions*, Trans. Amer. Math. Soc., 347 (1995), pp. 1713–1723.
[5] G.D. Anderson, M.K. Vamanamurthy, and M. Vuorinen, *Conformal Invariants, Inequalities, and Quasiconformal Mappings*, John Wiley & Sons, New York, 1997.
[6] G.D. Anderson, S.-L. Qiu, and M. Vuorinen, *Precise estimates for differences of the Gaussian hypergeometric function*, J. Math. Anal. Appl., 215 (1997), pp. 212–234.
[7] R. Askey, G. Gasper, and M. Ismail, *A positive sum from summability theory*, J. Approx. Theory, 13 (1975), pp. 413–420.
[8] R.W. Barnard, *On applications of hypergeometric functions*, J. Comput. Appl. Math., 105 (1999), pp. 1–8.
[9] R.W. Barnard, K. Pearce, and K.C. Richards, *An inequality involving the generalized hypergeometric function and the arc length of an ellipse*, SIAM J. Math. Anal., 31(2000), pp. 693–699.
[10] J.M. Borwein and P.B. Borwein, *Inequalities for compound mean iterations with logarithmic asymptotes*, J. Math. Anal. Appl., 177 (1993), pp. 572–582.
[11] G. Gasper and M. Rahman, *Basic Hypergeometric Series*, Cambridge University Press, Cambridge, UK, 1990.
[12] C. Maclaurin, *A Treatise of Fluxions in Two Books*, Vol. 2, T.W. and T. Ruddimans, Edinburgh, 1742.
[13] A.P. Prudnikov, Yu. A. Brychkov, and O.I. Marichev, *Integrals and Series, Vol.* 3: *More Special Functions*, Gordon and Breach Science Publishers, New York, 1990.
[14] E.D. Rainville, *Special Functions*, Macmillan, New York, 1960.
[15] M.K. Vamanamurthy and M. Vuorinen, *Inequalities for means*, J. Math. Anal. Appl., 183 (1994), pp. 155–166.
[16] M. Vuorinen, *Hypergeometric functions in geometric function theory*, in Proceedings of Special Functions and Differential Equations, K.S. Rao, R. Jagannathan, G. Vanden Berghe, and J. Van der Jeugt, eds., Allied Publishers, New Delhi, 1998.

SIAM J. MATH. ANAL.
Vol. 32, No. 2, pp. 420–434

DISTRIBUTIONAL SOLUTIONS OF NONHOMOGENEOUS DISCRETE AND CONTINUOUS REFINEMENT EQUATIONS*

RONG-QING JIA†, QINGTANG JIANG‡, AND ZUOWEI SHEN§

Abstract. Discrete and continuous refinement equations have been widely studied in the literature for the last few years, due to their applications to the areas of wavelet analysis and geometric modeling. However, there is no "universal" theorem that deals with the problem about the existence of compactly supported distributional solutions for both discrete and continuous refinement equations simultaneously. In this paper, we provide a uniform treatment for both equations. In particular, a complete characterization of the existence of distributional solutions of nonhomogeneous discrete and continuous refinement equations is given, which covers all cases of interest.

Key words. nonhomogeneous, discrete and continuous refinement equations, existence, uniqueness

AMS subject classifications. 41A58, 42C15, 41A17, 42C99

PII. S0036141099350882

1. Introduction and notations.

Let M be a dilation matrix, that is, an $s \times s$ real matrix whose eigenvalues lie outside the closed unit disk. We are interested in the following nonhomogeneous refinement equation:

$$\phi(x) = g(x) + \int_{\mathbb{R}^s} d\mu(y)\phi(Mx - y), \quad x \in \mathbb{R}^s, \tag{1.1}$$

where $\phi = (\phi_1, \ldots, \phi_r)^T$ is the unknown, $g = (g_1, \ldots, g_r)^T$ is a given $r \times 1$ vector of compactly supported distributions on $\mathbb{R}^s$, and μ is an $r \times r$ matrix of finite complex Borel measures on $\mathbb{R}^s$ with compact supports. Let $\mu = (\mu_{lj})_{1 \le l,j \le r}$. Then (1.1) can be written in the component form

$$\phi_l(x) = g_l(x) + \sum_{j=1}^{r} \int_{\mathbb{R}^s} \phi_j(Mx - y) d\mu_{lj}(y), \quad l = 1, \ldots, r. \tag{1.2}$$

When each μ_{lj} is a discrete Borel measure, (1.1) becomes a discrete refinement equation; when each μ_{lj} is absolutely continuous with respect to the Lebesgue measure, (1.1) is a continuous refinement equation. If $g = 0$, then (1.1) becomes a homogeneous refinement equation:

$$\phi(x) = \int_{\mathbb{R}^s} d\mu(y)\phi(Mx - y). \tag{1.3}$$

*Received by the editors January 25, 1999; accepted for publication (in revised form) February 21, 2000; published electronically July 11, 2000. The research is supported in part by NSERC Canada under Grant OGP 121336 and WSRP, NUS, under a grant from NSTB and MOE, Singapore.
http://www.siam.org/journals/sima/32-2/35088.html
†Department of Mathematical Sciences, University of Alberta, Edmonton, Canada T6G 2G1 (jia@xihu.math.ualberta.ca).
‡Department of Mathematics, National University of Singapore, 10 Kent Ridge Crescent, Singapore 119260. Current address: Department of Mathematical Sciences, University of Alberta, Edmonton, Canada T6G 2G1 (qjiang@haar.math.nus.edu.sg, jiang@math.ualberta.ca).
§Department of Mathematics, National University of Singapore, 10 Kent Ridge Crescent, Singapore 119260 (matzuows@leonis.nus.edu.sg).

Refinement equations are fundamental to wavelet theory and subdivision. In the context of wavelet theory, the key step to the construction of wavelets is to construct suitable refinable functions. In the context of subdivision, the limiting surface of a subdivision process is a linear combination of integer translates of the refinable function corresponding to the subdivision scheme.

For the scalar case ($r = 1$), a homogeneous discrete refinement equation can be written as

$$\phi(x) = \sum_{\alpha \in \mathbb{Z}^s} a(\alpha)\phi(Mx - \alpha), \quad x \in \mathbb{R}^s, \tag{1.4}$$

where the refinement mask a is finitely supported. Existence and uniqueness of the solutions of (1.4) were studied in [3] and [7] for the case when the dilation matrix M is two times the $s \times s$ identity matrix I_s. In particular, for the univariate case $s = 1$, it was proved in [7] that (1.4) has a nontrivial L_1-solution with compact support only if $\sum_{\alpha \in \mathbb{Z}^s} a(\alpha) = 2^n$ for some positive integer n.

For the vector case ($r > 1$), the coefficients $a(\alpha)$, $\alpha \in \mathbb{Z}^s$ in (1.4) become $r \times r$ complex matrices. Existence of compactly supported distributional solutions is characterized by spectral properties of the matrix $\Delta := \sum_\alpha a(\alpha)/|\det(M)|$. The spectral radius of Δ is denoted by $\rho(\Delta)$.

Existence (and uniqueness) of compactly supported distributional solutions of the vector refinement equation was investigated in [14] for the case when $s = 1$ and $M = (2)$. One of the main results of [14] states as follows: Suppose that there is a single eigenvalue λ of Δ with $|\lambda| = \rho(\Delta) < 2$; then the vector refinement equation (1.4) has k independent compactly supported distributional solutions, where k is the multiplicity of the eigenvalue 1 of Δ. This result was improved in [5]. It is still valid under a weak assumption that $\rho(\Delta) < 2$. A complete characterization of the existence of the compactly supported distributional solutions was given in [17] (for the case $M = 2I_s$) and in [24] (for the case $r = 2$, $s = 1$, and $M = (2)$). It states that the vector refinement equation (1.4) has a nontrivial compactly supported distributional solution if and only if there exists a nonnegative integer n such that 2^n is an eigenvalue for Δ.

Nonhomogeneous discrete refinement equations were investigated in [9] and [22]. For the case $s = 1, M = (2)$, and $r = 1$, necessary and sufficient conditions for existence and uniqueness of nontrivial compactly supported distributional solutions were given independently in [9] and [22].

Homogeneous continuous refinement equations were studied by many authors (see [4], [6], [8], [12], [15], [16], [18], [19], and [21]). The interested readers should consult the aforementioned references for details.

Although a lot of work has been done on this subject, there is no "universal" theorem that covers all cases. In this paper, we give a uniform treatment of the existence and uniqueness of distributional solutions of both discrete and continuous nonhomogeneous refinement equations in the most general setting, for the case of an arbitrary dilation matrix, any number of functions and any number of variables. The main idea is to use an iteration scheme in the Fourier domain with *real* variables. This approach enables us to unify the treatment for both discrete and continuous refinement equations.

While revising this paper, we became aware of recent papers of [10] and [23] related to our work. In contrast to our general results which are applicable to arbitrary dilation matrices, both papers deal with the case $M = 2I_s$ only.

Here is a brief outline of the present paper. Section 2 is devoted to a complete characterization of the existence of compactly supported distributional solutions of (1.1) in terms of g and μ. In section 3, several examples are given to illustrate our theory.

We now turn to the basics needed in this paper.

Let $\mathbb{C}^r$ denote the linear space of all $r \times 1$ complex vectors. The norm (or length) of a vector $\xi = (\xi_1, \ldots, \xi_r) \in \mathbb{C}^r$ is defined as

$$|\xi| := |\xi_1| + \cdots + |\xi_r|, \quad \xi = (\xi_1, \ldots, \xi_r) \in \mathbb{C}^r. \tag{1.5}$$

By $\mathbb{C}^{r\times r}$ we denote the linear space of all $r \times r$ complex matrices. For an $r \times r$ complex matrix $A = (a_{ij})_{1\le i,j\le r} \in \mathbb{C}^{r\times r}$, its norm is defined to be the maximum of the norm of its column vectors, i.e.,

$$\|A\| := \max\left\{ \sum_{i=1}^{r} |a_{ij}| : j = 1, 2, \ldots, r \right\}.$$

For a linear space F, F^r is denoted as the linear space

$$\left\{ (f_1, \ldots, f_r)^T : f_1, \ldots, f_r \in F \right\}.$$

When F is a Banach space equipped with the norm $\|\cdot\|$, the space F^r is also a Banach space with the norm given by

$$\|f\| := \sum_{j=1}^{r} \|f_j\|, \qquad f = (f_1, \ldots, f_r)^T \in F^r.$$

The space $\mathbb{R}^s$ is the s-dimensional Euclidean space equipped with the norm in (1.5). The set of all positive integers is denoted by $\mathbb{N}$; and $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$ is the set of all nonnegative integers.

A nonnegative integer $\alpha = (\alpha_1, \ldots, \alpha_s) \in \mathbb{N}_0^s$ is also used as a multi-index. Its length is the norm of α given in (1.5). For two multi-indices $\alpha = (\alpha_1, \ldots, \alpha_s)$ and $\beta = (\beta_1, \ldots, \beta_s)$, $\beta \le \alpha$ whenever $\beta_j \le \alpha_j$ for $j = 1, \ldots, s$.

For $\alpha = (\alpha_1, \ldots, \alpha_s) \in \mathbb{N}_0^s$ and $x = (x_1, \ldots, x_s) \in \mathbb{R}^s$, set $x^\alpha := x_1^{\alpha_1} \cdots x_s^{\alpha_s}$. We also use x^α to denote the function whose value at any x is x^α. The space P_n is the set of all polynomials of (total) degree at most n. For $j = 1, \ldots, s$, D_j denotes the partial derivative with respect to the jth coordinate and D^α is the differential operator $D_1^{\alpha_1} \cdots D_s^{\alpha_s}$. More generally, for a given polynomial $p(x) = \sum_\alpha c_\alpha x^\alpha$, $x \in \mathbb{R}^s$, the corresponding differential operator is

$$p(D) := \sum_\alpha c_\alpha D^\alpha.$$

Finally, for a given nonnegative integer α, the factorial of α is defined as $\alpha! := \alpha_1! \cdots \alpha_s!$.

Next, we list some basic notations of tempered distributions used in this paper. Let φ be a C^∞ function on $\mathbb{R}^s$. The seminorm $\|\cdot\|_{(m,\alpha)}$ of φ for a nonnegative integer m and a multi-index α is defined as

$$\|\varphi\|_{(m,\alpha)} := \sup_{x\in\mathbb{R}^s} \{(1 + |x|)^m |D^\alpha \varphi(x)|\}.$$

A function $\varphi \in C^\infty(\mathbb{R}^s)$ is said to be rapidly decreasing if $\|\varphi\|_{(m,\alpha)} < \infty \ \forall\ m \in \mathbb{N}_0$ and all $\alpha \in \mathbb{N}_0^s$. On the other hand, a continuous function f on $\mathbb{R}^s$ is said to be slowly increasing if there exists a polynomial p in s variables such that

$$|f(x)| \le |p(x)| \qquad \forall\, x \in \mathbb{R}^s.$$

Let $\mathcal{S}(\mathbb{R}^s)$ be the Schwartz space which is the space of all rapidly decreasing functions on $\mathbb{R}^s$ equipped with the metric

$$d(f,g) := \sum_{m=0}^{\infty} \sum_{|\alpha|=m} \frac{1}{2^m} \frac{\|f-g\|_{(m,\alpha)}}{1+\|f-g\|_{(m,\alpha)}}, \qquad f, g \in \mathcal{S}(\mathbb{R}^s).$$

A linear continuous functional on $\mathcal{S}(\mathbb{R}^s)$ is called a tempered distribution. The space $\mathcal{S}'(\mathbb{R}^s)$ is the linear space of all tempered distributions on $\mathbb{R}^s$. For example, the Dirac function δ given by

$$\langle \delta, \varphi \rangle := \varphi(0) \qquad \forall\, \varphi \in \mathcal{S}(\mathbb{R}^s)$$

is a tempered distribution. A slowly increasing continuous function $f \in \mathbb{R}^s$ induces a tempered distribution by

$$\langle f, \varphi \rangle := \int_{\mathbb{R}^s} f(x)\varphi(x)dx \qquad \forall\, \varphi \in \mathcal{S}(\mathbb{R}^s).$$

Let f be a tempered distribution on $\mathbb{R}^s$. We say that f vanishes on an open set $G \in \mathbb{R}^s$ if $\langle f, \varphi \rangle = 0$ for every $\varphi \in \mathcal{S}(\mathbb{R}^s)$ supported in G. Let W be the union of all open subsets G of $\mathbb{R}^s$ in which f vanishes. The complement of W is the support of f and denoted by $\mathrm{supp} f$. If $\mathrm{supp} f$ is a compact subset of $\mathbb{R}^s$, then we say that f is compactly supported.

The Fourier transform of a function φ in $\mathcal{S}(\mathbb{R}^s)$ is defined by

$$\widehat{\varphi}(\omega) := \int_{\mathbb{R}^s} \varphi(x) e^{-ix\cdot\omega} dx, \qquad \omega \in \mathbb{R}^s,$$

where i stands for the imaginary unit, and $x \cdot \omega := x_1\omega_1 + \cdots + x_s\omega_s$ for $x = (x_1, \ldots, x_s)$ and $\omega = (\omega_1, \ldots, \omega_s)$.

The Fourier transform of $f \in \mathcal{S}'(\mathbb{R}^s)$ is the tempered distribution $\widehat{f}$ defined by

$$\langle \widehat{f}, \varphi \rangle = \langle f, \widehat{\varphi} \rangle \qquad \forall\, \varphi \in \mathcal{S}(\mathbb{R}^s).$$

For example, the Fourier transform of the Dirac function δ is the constant 1. Let p be a polynomial and $f \in \mathcal{S}'(\mathbb{R}^s)$; the Fourier transform of $p(-iD)f$ is $p\widehat{f}$. In particular, the Fourier transform of $p(-iD)\delta$ is p.

The Fourier transform of a compactly supported distribution is an analytic function. Recall that a function f on $\mathbb{R}^s$ is said to be analytic if f can be expanded into a power series

$$f(x) = \sum_{\alpha \in \mathbb{N}_0^s} c_\alpha x^\alpha,$$

which converges for every $x \in \mathbb{R}^s$. The coefficients c_α are given by $c_\alpha = D^\alpha f(0)/\alpha!$. We use $\mathcal{A}(\mathbb{R}^s)$ to denote the linear space of all analytic functions on $\mathbb{R}^s$.

We will also use the following identity:

$$\langle f, \varphi\rangle = (2\pi)^{-s/2}\langle \widehat{f}, \widehat{\varphi}\rangle \qquad \forall\, f \in \mathcal{S}'(\mathbb{R}^s), \varphi \in \mathcal{S}(\mathbb{R}^s).$$

A vector of distributions $\phi = (\phi_1, \ldots, \phi_r)^T \in (\mathcal{S}'(\mathbb{R}^s))^r$ is called a solution of (1.1) if

$$\langle \phi, \varphi\rangle = \langle g, \varphi\rangle + \left\langle \phi, \int_{\mathbb{R}^s} \overline{(d\mu(y))^T}\varphi(M^{-1}(\cdot + y))/|\det(M)| \right\rangle$$

holds $\forall \varphi = (\varphi_1, \ldots, \varphi_r)^T \in (\mathcal{S}(\mathbb{R}^s))^r$.

2. Existence of solutions. The problem of the existence of distributional solutions of discrete and continuous refinement equations can be handled simultaneously in the Fourier domain. For this, recall that the Fourier transform of μ_{lj} is given by

$$\widehat{\mu}_{lj}(\omega) = \int_{\mathbb{R}^s} e^{-i\omega\cdot y}\, d\mu_{lj}(y), \quad \omega \in \mathbb{R}^s.$$

Thus, refinement equation (1.1) can be written as

$$\widehat{\phi}(\omega) = \widehat{g}(\omega) + H(N\omega)\widehat{\phi}(N\omega), \quad \omega \in \mathbb{R}^s, \tag{2.1}$$

where $N := (M^{-1})^T$ and

$$H(\omega) := (1/|\det(M)|)\widehat{\mu}(\omega) = (1/|\det(M)|)(\widehat{\mu}_{lj}(\omega))_{1\le l,j\le r}, \quad \omega \in \mathbb{R}^s. \tag{2.2}$$

For a nonnegative integer n, we denote by P_n^r the linear space of all $r \times 1$ vectors of polynomials of degree at most n. For $f \in (\mathcal{A}(\mathbb{R}^s))^r$, define

$$f^{[n]}(\omega) := \sum_{|\alpha|\le n} D^\alpha f(0)\omega^\alpha/\alpha!, \qquad \omega \in \mathbb{R}^s.$$

Clearly, $f^{[n]}$ belongs to P_n^r. Let L_n be the linear operator defined on P_n^r by

$$L_n p := \big(H(N\cdot)p(N\cdot)\big)^{[n]}, \quad p \in P_n^r.$$

The linear operator L_n can be viewed as follows. Let $\sum_{\alpha\in\mathbb{N}_0^s} v_\alpha\omega^\alpha$, $\omega \in \mathbb{R}^s$, be the Taylor expansion of $H(N\omega)p(N\omega)$. Then, $L_n p(\omega) = \sum_{|\alpha|\le n} v_\alpha\omega^\alpha$.

Suppose $\widehat{\phi}$ satisfies (2.1). Then for any $n \in \mathbb{N}_0$,

$$\widehat{\phi}^{[n]} = \widehat{g}^{[n]} + \big(H(N\cdot)\widehat{\phi}(N\cdot)\big)^{[n]} = \widehat{g}^{[n]} + L_n\widehat{\phi}^{[n]}.$$

Hence, $p := \widehat{\phi}^{[n]} \in P_n^r$ is a solution of the following linear equation:

$$p - L_n p = \widehat{g}^{[n]}. \tag{2.3}$$

Next we show that if (2.3) has a solution $p \in P_n^r$ for a sufficiently large integer n, then (2.1) has a compactly supported distributional solution ϕ such that $\widehat{\phi}^{[n]} = p$. For this, we first note that if f is a compactly supported *continuous function*, then using Taylor's formula (see, e.g., [20, Theorem 7.7]) we have

$$\widehat{f}(\omega) = \sum_{|\alpha|\le n} \frac{D^\alpha\widehat{f}(0)}{\alpha!}\omega^\alpha + \sum_{|\alpha|=n+1} \frac{D^\alpha\widehat{f}(\xi)}{\alpha!}\omega^\alpha,$$

where ξ is a point on the straight line segment from 0 to ω. Note that

$$D^\alpha \widehat{f}(\xi) = \int_{\mathbb{R}^s} (-ix)^\alpha f(x) e^{-ix\cdot\xi}\, dx.$$

Since f is a compactly supported continuous function, the set $K := \operatorname{supp} f$ is a compact set of $\mathbb{R}^s$. Hence,

$$\left|D^\alpha \widehat{f}(\xi)\right| \le \int_K |x|^{|\alpha|} |f(x)|\, dx < \infty.$$

Therefore, there exists a constant C_n such that

$$\left| \widehat{f}(\omega) - \sum_{|\alpha|\le n} D^\alpha \widehat{f}(0)\omega^\alpha/\alpha! \right| \le C_n |\omega|^{n+1} \qquad \forall\, \omega \in \mathbb{R}^s. \tag{2.4}$$

The following lemma extends the above estimate to compactly supported *distributions*. The key to our extension is the well-known fact that a *compactly supported* distribution is of finite order (see [2, Theorem 2.22]).

LEMMA 2.1. *Suppose f is a compactly supported distribution on $\mathbb{R}^s$. Then for a given nonnegative integer n, there exists a polynomial q_n in s variables such that*

$$\left| \widehat{f}(\omega) - \sum_{|\alpha|\le n} D^\alpha \widehat{f}(0)\omega^\alpha/\alpha! \right| \le |\omega|^{n+1} q_n(\omega) \qquad \forall\, \omega \in \mathbb{R}^s. \tag{2.5}$$

Proof. Since f is compactly supported, there exists a positive integer m and compactly supported continuous functions f_β ($|\beta| \le m$) such that $f = \sum_{|\beta|\le m} D^\beta f_\beta$. Hence,

$$\widehat{f}(\omega) = \sum_{|\beta|\le m} (i\omega)^\beta \widehat{f}_\beta(\omega), \qquad \omega \in \mathbb{R}^s.$$

Set $c_{\alpha,\beta} := D^\alpha \widehat{f}_\beta(0)/\alpha!$ for $\alpha \in \mathbb{N}_0^s$ and $|\beta| \le m$. Write $\widehat{f}$ as the sum of h_1 and h_2, where

$$h_1(\omega) := \sum_{|\beta|\le m} (i\omega)^\beta \sum_{|\alpha|\le n} c_{\alpha,\beta}\omega^\alpha, \qquad \omega \in \mathbb{R}^s,$$

and

$$h_2(\omega) := \sum_{|\beta|\le m} (i\omega)^\beta \left(\widehat{f}_\beta(\omega) - \sum_{|\alpha|\le n} c_{\alpha,\beta}\omega^\alpha \right), \qquad \omega \in \mathbb{R}^s.$$

It follows from (2.4) that there exists a polynomial u such that

$$|h_2(\omega)| \le |\omega|^{n+1} u(\omega) \qquad \forall\, \omega \in \mathbb{R}^s.$$

But h_1 is a polynomial, so there exists a polynomial v such that

$$\left| h_1(\omega) - \sum_{|\alpha|\le n} D^\alpha h_1(0)\omega^\alpha/\alpha! \right| \le |\omega|^{n+1} v(\omega) \qquad \forall\, \omega \in \mathbb{R}^s.$$

Since $D^\alpha \widehat{f}(0) = D^\alpha h_1(0)\ \forall\ |\alpha| \le n$, we conclude that (2.5) holds with $q_n = u + v$. ∎

To state the next theorem, we define

$$c_0 := \sup_{\omega\in\mathbb{R}^s} \|H(\omega)\|. \tag{2.6}$$

Since each measure μ_{lj} $(l, j = 1, \dots, r)$ is finite, by (2.2), $\|H(\omega)\|$ is bounded on $\mathbb{R}^s$. Hence, $c_0 < \infty$. We also recall that $\rho(N)$ is the spectral radius of the matrix N.

THEOREM 2.2. *Suppose* (2.3) *has a solution* $p \in P_n^r$ *for some nonnegative integer* n *satisfying* $\rho(N)^{n+1} < 1/c_0$. *Then* (2.1) *has a compactly supported distributional solution* ϕ *such that* $\widehat{\phi}^{[n]} = p$.

Proof. In this proof, the number n is fixed. The proof is based on the following iteration scheme. It starts with the $r \times 1$ vector $\phi_0 := p(-iD)\delta$, with the jth entry of ϕ_0 $p_j(-iD)\delta$, where p_j is the jth entry of the vector p and δ is the Dirac function. Each entry of ϕ_0 is supported at the origin and $\widehat{\phi}_0 = p$. For $k = 1, 2, \dots$, the $r \times 1$ vectors ϕ_k are defined recursively by

$$\widehat{\phi}_k(\omega) := \widehat{g}(\omega) + H(N\omega)\widehat{\phi}_{k-1}(N\omega). \tag{2.7}$$

In particular, $\widehat{\phi}_1^{[n]} = \widehat{g}^{[n]} + L_n p = p$. By (2.7) we have

$$\begin{aligned}\widehat{\phi}_{k+1}(\omega) - \widehat{\phi}_k(\omega) &= H(N\omega)\big(\widehat{\phi}_k(N\omega) - \widehat{\phi}_{k-1}(N\omega)\big)\\ &= \left(\prod_{j=1}^{k} H(N^j\omega)\right)\big(\widehat{\phi}_1(N^k\omega) - \widehat{\phi}_0(N^k\omega)\big).\end{aligned}$$

Since $\|H(\omega)\| \le c_0\ \forall\ \omega \in \mathbb{R}^s$, we have

$$|\widehat{\phi}_{k+1}(\omega) - \widehat{\phi}_k(\omega)| \le c_0^k |\widehat{\phi}_1(N^k\omega) - \widehat{\phi}_0(N^k\omega)| \qquad \forall\, \omega \in \mathbb{R}^s \text{ and } k \in \mathbb{N}_0. \tag{2.8}$$

Note that $\widehat{\phi}_1^{[n]} - \widehat{\phi}_0^{[n]} = 0$. By Lemma 2.1, there exists a polynomial q (depending on n) such that

$$|\widehat{\phi}_1(N^k\omega) - \widehat{\phi}_0(N^k\omega)| \le |N^k\omega|^{n+1} q(N^k\omega) \qquad \forall\, \omega \in \mathbb{R}^s \text{ and } k \in \mathbb{N}_0. \tag{2.9}$$

Since $\rho(N)^{n+1} < 1/c_0$ and $\rho(N) < 1$, there is $\varepsilon > 0$ such that $t := (\rho(N)+\varepsilon)^{n+1} c_0 < 1$ and $\rho(N) + \varepsilon < 1$. Hence,

$$|N^k\omega| \le C(\rho(N) + \varepsilon)^k |\omega| \qquad \forall\, \omega \in \mathbb{R}^s \text{ and } k \in \mathbb{N}_0, \tag{2.10}$$

for some constant C (depending on ε). This implies that there exists a polynomial Q with $q(N^k\omega) \le Q(\omega)\ \forall\ k \in \mathbb{N}_0$ and all $\omega \in \mathbb{R}^s$. Combining (2.8), (2.9), and (2.10), we obtain

$$|\widehat{\phi}_{k+1}(\omega) - \widehat{\phi}_k(\omega)| \le t^k |C\omega|^{n+1} Q(\omega) \qquad \forall\, \omega \in \mathbb{R}^s \text{ and } k \in \mathbb{N}_0, \tag{2.11}$$

which means that for each $\omega \in \mathbb{R}^s$, $(\widehat{\phi}_k(\omega))_{k\in\mathbb{N}}$ is a Cauchy sequence. Hence,

$$f(\omega) := \lim_{k\to\infty} \widehat{\phi}_k(\omega), \quad \omega \in \mathbb{R}^s,$$

is well defined. Moreover, $(\widehat{\phi}_k)_{k\in\mathbb{N}}$ converges to f uniformly on an arbitrary compact subset of $\mathbb{R}^s$. So f is an $r\times 1$ vector of continuous functions on $\mathbb{R}^s$. Furthermore, we deduce from (2.11) that

$$\begin{aligned}|\widehat{\phi}_k(\omega)-p(\omega)| &\le \sum_{j=0}^{k-1}|\widehat{\phi}_{j+1}(\omega)-\widehat{\phi}_j(\omega)|\\ &\le (1-t)^{-1}|C\omega|^{n+1}Q(\omega),\quad \omega\in\mathbb{R}^s.\end{aligned}\tag{2.12}$$

Consequently,

$$|f(\omega)-p(\omega)|\le (1-t)^{-1}|C\omega|^{n+1}Q(\omega),\quad \omega\in\mathbb{R}^s.$$

Hence, f is an $r\times 1$ vector of slowly increasing continuous functions with $f^{[n]}=p$. Therefore, there is a unique $\phi\in(\mathcal{S}'(\mathbb{R}^s))^r$ such that $f=\widehat{\phi}$, and ϕ satisfies (2.1).

It remains to prove that ϕ is compactly supported. Let K be a compact subset of $\mathbb{R}^s$ such that

$$\{0\}\cup \operatorname{supp}\,\mu\cup\big(M(\operatorname{supp}\,g)\big)\subseteq K.$$

Let

$$\Omega:=\sum_{n=1}^{\infty}M^{-n}K.$$

Recall that $\phi_0=p(-iD)\delta$. By our choice of K,

$$\operatorname{supp}\,\phi_0=\{0\}\subseteq K.$$

It can be easily proved inductively that $\operatorname{supp}\phi_k\subseteq\Omega\ \forall\ k\in\mathbb{N}_0$ (see [13]). Suppose φ belongs to $(\mathcal{S}(\mathbb{R}^s))^r$ and $\operatorname{supp}\varphi\subset\mathbb{R}^s\setminus\Omega$. Since $\widehat{\varphi}$ is rapidly decreasing and since (2.12) is valid, there exists a constant C such that

$$\big|\widehat{\phi}_k(\omega)^T\widehat{\varphi}(\omega)\big|\le C(1+|\omega|)^{-s-1}\qquad\forall\,\omega\in\mathbb{R}^s\text{ and }k\in\mathbb{N}_0.$$

Thus, the Lebesgue dominated convergence theorem leads to

$$\lim_{k\to\infty}\int_{\mathbb{R}^s}\widehat{\phi}_k(\omega)^T\widehat{\varphi}(\omega)\,d\omega=\int_{\mathbb{R}^s}f(\omega)^T\widehat{\varphi}(\omega)\,d\omega.$$

In other words, $\lim_{k\to\infty}\langle\widehat{\phi}_k,\widehat{\varphi}\rangle=\langle\widehat{\phi},\widehat{\varphi}\rangle$. Therefore, we obtain

$$\langle\phi,\varphi\rangle=(2\pi)^{-s/2}\langle\widehat{\phi},\widehat{\varphi}\rangle=(2\pi)^{-s/2}\lim_{k\to\infty}\langle\widehat{\phi}_k,\widehat{\varphi}\rangle=\lim_{k\to\infty}\langle\phi_k,\varphi\rangle=0.$$

Hence, $\langle\phi,\varphi\rangle=0\ \forall\ \varphi\in(\mathcal{S}(\mathbb{R}^s))^r$ supported in $\mathbb{R}^s\setminus\Omega$, which implies that ϕ is supported in Ω. □

Theorem 2.2 reduces the problem of the existence of solutions of (1.1) to that of (2.3).

In order to study (2.3), we shall use the notation introduced in [1]. For $|\beta|=k$, write

$$(M^T\omega)^\beta=\sum_{|\alpha|=k}m_{\alpha,\beta}\omega^\alpha,\quad\omega\in\mathbb{R}^s.$$

The coefficients $m_{\alpha,\beta}$ ($|\alpha| = k, |\beta| = k$) are uniquely determined by the matrix M and the number k. The matrix $(m_{\alpha,\beta})_{|\alpha|=k,|\beta|=k}$ will be denoted by M_k. For $k \in \mathbb{N}_0$, let J_k be the set $\{\alpha \in \mathbb{N}_0^s : |\alpha| = k\}$. The cardinality of J_k is $d_k := \binom{k+s-1}{s-1}$. The ordering $\prec$ on J_k is defined as follows. For $\alpha = (\alpha_1, \ldots, \alpha_s) \in J_k$ and $\beta = (\beta_1, \ldots, \beta_s) \in J_k$, $\alpha \prec \beta$ whenever there exists some j, $1 \le j \le s$, such that $\alpha_j < \beta_j$ and $\alpha_i = \beta_i$ for $i = j+1, \ldots, s$.

Replacing ω by $M^T\omega$ in (2.1), we have

$$\hat{\phi}(M^T\omega) = \hat{g}(M^T\omega) + H(\omega)\hat{\phi}(\omega), \quad \omega \in \mathbb{R}^s. \tag{2.13}$$

Write $\hat{\phi}(\omega) = \sum_{\beta\in\mathbb{N}_0^s} v_\beta \omega^\beta$, $H(\omega) = \sum_{\beta\in\mathbb{N}_0^s} H_\beta \omega^\beta$, and $\hat{g}(\omega) = \sum_{\beta\in\mathbb{N}_0^s} g_\beta \omega^\beta$, $\omega \in \mathbb{R}^s$. Substituting the above expressions into (2.13) and comparing the coefficients of both sides, we obtain

$$\sum_{|\beta|=k} m_{\alpha,\beta} v_\beta - \sum_{0\le\gamma\le\alpha} H_{\alpha-\gamma} v_\gamma = h_\alpha, \qquad |\alpha| = k, \tag{2.14}$$

where

$$h_\alpha := \sum_{|\beta|=k} m_{\alpha,\beta} g_\beta, \qquad |\alpha| = k. \tag{2.15}$$

Denote by $v_{[k]}$ the $(rd_k) \times 1$ column vector defined by $v_{[k]} := (v_\beta)_{|\beta|=k}$. The column vector $v_{[k]}$ is ordered from the top to the bottom as follows. For α, β with $|\alpha| = |\beta| = k$, if $\alpha \prec \beta$, then the segment v_α is put at the top of the segment v_β. The $(rd_k) \times 1$ column vector $h_{[k]} := (h_\beta)_{|\beta|=k}$ is defined similarly.

The notation $B \otimes C$ stands for $(b_{ij}C)$, the (right) Kronecker product of two matrices $B = (b_{ij})$ and C. With this, (2.14) can be rewritten as

$$(M_k \otimes I_r)\, v_{[k]} - \sum_{j=0}^{k} (H_{\alpha-\gamma})_{|\alpha|=k,|\gamma|=j} v_{[j]} = h_{[k]}, \tag{2.16}$$

where $H_{\alpha-\gamma}$ is understood to be 0 if $\gamma \le \alpha$ does not hold. When $|\alpha| = k$ and $|\gamma| = k$, we have $(H_{\alpha-\gamma})_{|\alpha|=k,|\gamma|=k} = I_{d_k} \otimes H(0)$. It follows from (2.16) that

$$T_k \begin{bmatrix} v_{[0]} \\ v_{[1]} \\ \vdots \\ v_{[k]} \end{bmatrix} = \begin{bmatrix} h_{[0]} \\ h_{[1]} \\ \vdots \\ h_{[k]} \end{bmatrix}, \tag{2.17}$$

where the matrix T_k is given by

$$T_k := \begin{bmatrix} I_r & 0 & 0 & \cdots & 0 \\ 0 & M_1 \otimes I_r & 0 & \cdots & 0 \\ 0 & 0 & M_2 \otimes I_r & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & M_k \otimes I_r \end{bmatrix} \tag{2.18}$$

$$- \begin{bmatrix} H(0) & 0 & 0 & \cdots & 0 \\ (H_\alpha)_{|\alpha|=1} & I_{d_1} \otimes H(0) & 0 & \cdots & 0 \\ (H_\alpha)_{|\alpha|=2} & (H_{\alpha-\gamma})_{|\alpha|=2,|\gamma|=1} & I_{d_2} \otimes H(0) & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ (H_\alpha)_{|\alpha|=k} & (H_{\alpha-\gamma})_{|\alpha|=k,|\gamma|=1} & (H_{\alpha-\gamma})_{|\alpha|=k,|\gamma|=2} & \cdots & I_{d_k} \otimes H(0) \end{bmatrix}.$$

Therefore, ϕ satisfies (2.1) if and only if, for each $k \in \mathbb{N}_0$, $v_{[0]}, v_{[1]}, \ldots, v_{[k]}$ satisfy (2.17).

Let $\lambda_1, \ldots, \lambda_s$ be the eigenvalues of M. As usual, for $\beta = (\beta_1, \ldots, \beta_s) \in \mathbb{N}_0^s$, $\lambda^\beta := \lambda_1^{\beta_1} \cdots \lambda_s^{\beta_s}$.

LEMMA 2.3. *Let k be a nonnegative integer. Suppose λ^β is not an eigenvalue of $H(0)$ for any $\beta \in \mathbb{N}_0^s$ with $|\beta| = k$. Then the matrix*

$$M_k \otimes I_r - I_{d_k} \otimes H(0)$$

is nonsingular.

Proof. Suppose B is an $s \times s$ matrix. For $\alpha \in J_k$ we write

$$(B\omega)^\alpha = \sum_{|\beta|=k} b_{\alpha,\beta}^{[k]} \omega^\beta, \quad \omega \in \mathbb{R}^s,$$

where $b_{\alpha,\beta}^{[k]}$ are complex numbers. Let $B^{[k]}$ denote the matrix $(b_{\alpha,\beta}^{[k]})_{|\alpha|=k,|\beta|=k}$. Suppose C is also an $s \times s$ matrix. It is easily seen that

$$(BC)^{[k]} = B^{[k]} C^{[k]}.$$

Since the eigenvalues of M are $\lambda_1, \ldots, \lambda_s$, there exists an invertible $s \times s$ matrix U such that the matrix $\Lambda := U^{-1} M^T U$ is a lower triangular matrix with $\lambda_1, \ldots, \lambda_s$ being the entries in its main diagonal. We also note that $M_k = ((M^T)^{[k]})^T$ by the definition of M_k.

In order to establish the lemma, it suffices to show that the matrix $(M^T)^{[k]} \otimes I_r - I_{d_k} \otimes H(0)^T$ is nonsingular. For that, we observe that

$$\left[(U^{-1})^{[k]} \otimes I_r\right] \left[(M^T)^{[k]} \otimes I_r - I_{d_k} \otimes H(0)^T\right] \left[U^{[k]} \otimes I_r\right] = \Lambda^{[k]} \otimes I_r - I_{d_k} \otimes H(0)^T.$$

Clearly, $\Lambda^{[k]}$ is a lower triangular matrix with λ^β ($|\beta| = k$) being the entries in its main diagonal. Thus, $\Lambda^{[k]} \otimes I_r - I_{d_k} \otimes H(0)^T$ is a lower triangular block matrix with diagonal blocks $\lambda^\beta I_r - H(0)^T$. Since λ^β is not an eigenvalue of $H(0)$ for any β with $|\beta| = k$, we conclude that the matrix $\Lambda^{[k]} \otimes I_r - I_{d_k} \otimes H(0)^T$ is nonsingular. □

We are in a position to establish the main result of this paper. In what follows, for $k \in \mathbb{N}_0$, T_k is the matrix given in (2.18), and $h_{[k]}$ is the vector $(h_\alpha)_{|\alpha|=k}$ with h_α given in (2.15). Finally, $\lambda_1, \ldots, \lambda_s$ are the eigenvalues of M.

THEOREM 2.4. *Suppose $H(0)$ has no eigenvalues of the form λ^β, $\beta \in \mathbb{N}_0^s$, then (1.1) has a unique compactly supported distributional solution. Suppose $H(0)$ has eigenvalues of the form λ^β for some $\beta \in \mathbb{N}_0^s$. Let $n_0 := \max\{|\beta| : \lambda^\beta$ is an eigenvalue of $H(0)\}$. Then (1.1) has a compactly supported distributional solution ϕ if and only if the linear equation*

$$T_{n_0} \begin{bmatrix} v_{[0]} \\ v_{[1]} \\ \vdots \\ v_{[n_0]} \end{bmatrix} = \begin{bmatrix} h_{[0]} \\ h_{[1]} \\ \vdots \\ h_{[n_0]} \end{bmatrix} \tag{2.19}$$

has a solution. Furthermore, let $v_{[0]}, v_{[1]}, \ldots, v_{[n_0]}$ be a solution of the above linear equation and $v_{[j]} = (v_\alpha)_{|\alpha|=j}$ ($j = 0, \ldots, n_0$). Then there is a unique compactly supported distributional solution ϕ of (1.1) satisfying $\widehat{\phi}^{[n_0]}(\omega) = \sum_{|\alpha| \le n_0} v_\alpha \omega^\alpha$, $\omega \in \mathbb{R}^s$.

Proof. Let n be a nonnegative integer satisfying $\rho(N)^{n+1} < 1/c_0$, where c_0 is given by (2.6). Suppose that $H(0)$ has no eigenvalues of the form λ^β, $\beta \in \mathbb{N}_0^s$. Then $M_j \otimes I_r - I_{d_j} \otimes H(0)$ is nonsingular for every $j \in \mathbb{N}_0$ by Lemma 2.3. Therefore, there is a unique solution $v_{[0]}, v_{[1]}, \ldots, v_{[n]}$ that satisfies the linear equation (2.17) for $k = n$. Hence, (1.1) has a unique compactly supported distributional solution by Theorem 2.2.

Next, suppose $H(0)$ has eigenvalues of the form λ^β for some $\beta \in \mathbb{N}_0^s$. Let $v_{[0]}, v_{[1]}, \ldots, v_{[n_0]}$ be a solution of the linear equation (2.19). By Lemma 2.3, $M_k \otimes I_r - I_{d_k} \otimes H(0)$ is nonsingular for $k > n_0$. Hence, we can find $v_{[n_0+1]}, \ldots, v_{[n]}$ from $v_{[0]}, \ldots, v_{[n_0]}$ by using (2.17) for $k = n_0 + 1, \ldots, n$. This implies that (2.3) has a solution $p = \sum_{|\beta| \le n} p_\beta \omega^\beta \in P_n^r$ with $(p_\beta)_{|\beta|=k} = v_{[k]}$, $0 \le k \le n$. By Theorem 2.2, (1.1) has a compactly supported distributional solution ϕ such that $\widehat{\phi}^{[n]} = p$. Consequently, $\widehat{\phi}^{[n_0]}(\omega) = \sum_{|\alpha| \le n_0} v_\alpha \omega^\alpha$, $\omega \in \mathbb{R}^s$, where v_α ($|\alpha| \le n_0$) are determined by $(v_\alpha)_{|\alpha|=j} = v_{[j]}$ for $j = 0, \ldots, n_0$.

Finally, we establish the uniqueness of the solution. Let ϕ and ψ be two compactly supported distributional solutions of (1.1) with $\hat{\phi}^{[n_0]} = \hat{\psi}^{[n_0]}$. Write $\hat{\phi}(\omega) = \sum_{\alpha \in \mathbb{N}_0^s} v_\alpha \omega^\alpha$ and $\hat{\psi}(\omega) = \sum_{\alpha \in \mathbb{N}_0^s} u_\alpha \omega^\alpha$, $\omega \in \mathbb{R}^s$. For $k \in \mathbb{N}_0$, let $v_{[k]} := (v_\alpha)_{|\alpha|=k}$ and $u_{[k]} := (u_\alpha)_{|\alpha|=k}$. We claim that $u_{[k]} = v_{[k]}$ $\forall\, k \in \mathbb{N}_0$. This is shown by induction on k. It is clear that $u_{[k]} = v_{[k]}$ for $k = 0, \ldots, n_0$. Consider $k > n_0$. Assume that $u_{[j]} = v_{[j]}$ for $j = 0, \ldots, k-1$. It follows from (2.17) that

$$\begin{aligned}
&(M_k \otimes I_r)\, u_{[k]} - \sum_{j=0}^{k} (H_{\alpha-\gamma})_{|\alpha|=k,|\gamma|=j} u_{[j]} \\
&= (M_k \otimes I_r)\, v_{[k]} - \sum_{j=0}^{k} (H_{\alpha-\gamma})_{|\alpha|=k,|\gamma|=j} v_{[j]}.
\end{aligned}$$

Since $u_{[j]} = v_{[j]}$ for $j = 0, \ldots, k-1$, we have that

$$(M_k \otimes I_r - I_{d_k} \otimes H(0))\, u_{[k]} = (M_k \otimes I_r - I_{d_k} \otimes H(0))\, v_{[k]}.$$

But the matrix $M_k \otimes I_r - I_{d_k} \otimes H(0)$ is nonsingular for $k > n_0$. Therefore, $u_{[k]} = v_{[k]}$. This shows $\phi = \psi$, as desired. ∎

When $H(0)$ has no eigenvalues of the form λ^β, $\beta \in \mathbb{N}_0^s$, the homogeneous equation (1.3) has only the trivial solution. The following corollary generalizes the result of [14], [17], and [24] to an arbitrary dilation matrix.

COROLLARY 2.5. *Homogeneous refinement equation* (1.3) *has a nontrivial compactly supported distributional solution if and only if* $H(0)$ *has an eigenvalue of the form* λ^β, $\beta \in \mathbb{N}_0^s$. *Furthermore, the number of linearly independent compactly supported solutions of* (1.3) *is the same as the dimension of the space* $\ker(T_{n_0})$, *where* $n_0 := \max\{|\beta| : \lambda^\beta$ *is an eigenvalue of* $H(0)\}$.

Suppose that ϕ and ψ are two compactly supported distributional solutions of (1.1). Then $\phi - \psi$ is a solution of the corresponding homogeneous equation (1.3). Thus we have the following corollary.

COROLLARY 2.6. *Suppose* $H(0)$ *has eigenvalues of the form* λ^β *for some* $\beta \in \mathbb{N}_0^s$. *Let* S *be the set of all compactly supported distributional solutions of* (1.1). *If* (1.1) *has at least one solution, then* S *is a linear manifold whose dimension is the same as that of* $\ker(T_{n_0})$, *where* $n_0 := \max\{|\beta| : \lambda^\beta$ *is an eigenvalue of* $H(0)\}$.

3. Examples. In this section we give several examples to illustrate our theory.

Example 1. The following interesting example was first studied in [11]. Let $r = 2$ and $s = 1$. Consider the discrete refinement equation

$$\phi = \sum_{j=0}^{3} a(j)\phi(2\cdot - j), \tag{3.1}$$

where

$$a(0) = \begin{bmatrix} h_1 & 1 \\ h_2 & h_3 \end{bmatrix}, \qquad a(1) = \begin{bmatrix} h_1 & 0 \\ h_4 & 1 \end{bmatrix},$$

$$a(2) = \begin{bmatrix} 0 & 0 \\ h_4 & h_3 \end{bmatrix}, \qquad a(3) = \begin{bmatrix} 0 & 0 \\ h_2 & 0 \end{bmatrix},$$

and h_1, h_2, h_3, h_4 are given by

$$h_1 = -\frac{t^2 - 4t - 3}{2(t+2)}, \qquad h_2 = -\frac{3(t^2-1)(t^2-3t-1)}{4(t+2)^2},$$

$$h_3 = \frac{3t^2 + t - 1}{2(t+2)}, \qquad h_4 = -\frac{3(t^2-1)(t^2-t+3)}{4(t+2)^2}.$$

It was proved in [11] that the refinement equation has a unique continuous nontrivial solution $\phi = (\phi_1, \phi_2)^T$ for $|t| < 1$. In particular, when $t = -0.2$, the shifts of ϕ_1 and ϕ_2 are orthogonal, and corresponding orthogonal double wavelets were constructed there.

Consider the case $|t| > 1$. We note that $H(\omega) = \sum_{j=0}^{3} a(j)e^{-ij\omega}/2$, $\omega \in \mathbb{R}$. The matrix $H(0)$ has two eigenvalues 1 and t. Therefore, (3.1) has compactly supported distributional solutions only if $t = 2^n$ for some positive integer (see [14]). The case $t = 2$ was discussed in [24], and it was shown there that (3.1) has *two* linearly independent solutions.

Here we consider the case $t = 4$. Write $H(\omega) = H_0 + H_1\omega + H_2\omega^2 + \cdots$, $\omega \in \mathbb{R}$, where $H_0, H_1, H_2, \ldots$ are 2×2 matrices. For the case $t = 4$, $n_0 = 2$, the corresponding matrix T_2 is given by

$$T_2 = -\begin{bmatrix} H(0) - I_2 & 0 & 0 \\ H_1 & H(0) - 2I_2 & 0 \\ H_2 & H_1 & H(0) - 4I_2 \end{bmatrix}.$$

A simple computation yields $\dim(\ker(T_2)) = 1$ for $t = 4$. By Corollary 2.5 we conclude that (3.1) has *one* linearly independent compactly supported distributional solution. Moreover, if ϕ is a nontrivial solution of (3.1), then we must have $\hat{\phi}(0) = 0$. This is in sharp contrast to the case $t = 2$.

Example 2. Let $r = 2$, $M = 2I_s$, and $g = 0$. Suppose

$$H(\omega) = \begin{bmatrix} h_{11}(\omega) & h_{12}(\omega) \\ h_{21}(\omega) & h_{22}(\omega) \end{bmatrix}, \quad \omega \in \mathbb{R}^s, \qquad \text{and} \qquad H(0) = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}. \tag{3.2}$$

In this case, $n_0 = 1$, and

$$T_1 = -\begin{bmatrix} H(0) - I_2 & & & & \\ D_1 H(0) & H(0) - 2I_2 & & & \\ D_2 H(0) & 0 & H(0) - 2I_2 & & \\ \vdots & \vdots & \vdots & \ddots & \\ D_s H(0) & 0 & 0 & \cdots & H(0) - 2I_2 \end{bmatrix}.$$

If $D_j h_{21}(0) = 0 \ \forall \ j = 1, \ldots, s$, then $\dim(\ker(T_1)) = s + 1$. So (1.3) has exactly $s + 1$ linearly independent solutions by Corollary 2.5. Otherwise, (1.3) has exactly s linearly independent solutions. Moreover, the homogeneous refinement equation (1.3) has a compactly supported distributional solution ϕ such that $\hat{\phi}(0) \neq 0$ if and only if $D_j h_{21}(0) = 0 \ \forall \ j = 1, \ldots, s$. For discrete refinement equations, this recovers the result of Theorem 4 in [24].

The next two examples are devoted to nonhomogeneous refinement equations.

Example 3. Let $r = 2$, $s = 1$, $M = (2)$, and $g = (g_1, g_2)^T$. Suppose the conditions in (3.2) are satisfied. In this case, $n_0 = 1$, and T_1 is the 4×4 matrix given by

$$T_1 = -\begin{bmatrix} H(0) - I_2 & 0 \\ H'(0) & H(0) - 2I_2 \end{bmatrix} = -\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ h'_{11}(0) & h'_{12}(0) & -1 & 0 \\ h'_{21}(0) & h'_{22}(0) & 0 & 0 \end{bmatrix}.$$

By Theorem 2.4, (1.1) has a compactly supported distributional solution if and only if the linear equation

$$T_1 v = [\hat{g}_1(0), \hat{g}_2(0), 2\hat{g}'_1(0), 2\hat{g}'_2(0)]^T$$

has a solution v in $\mathbb{C}^4$. Let S be the set of all compactly supported distributional solutions of (1.1). There are two possible cases: $h'_{21}(0) \neq 0$ and $h'_{21}(0) = 0$. In the former case, (1.1) has a solution if and only if $\hat{g}_1(0) = 0$, and $\dim(S) = 1$ by Corollary 2.6. In the latter case, i.e., $h'_{21}(0) = 0$, (1.1) has a compactly supported distributional solution if and only if $\hat{g}_1(0) = 0$ and $\hat{g}_2(0) h'_{22}(0) = 2\hat{g}'_2(0)$. If this is the case, then $\dim(S) = 2$.

Example 4. Let $r = 2$, $s = 2$, $M = 2I_2$, and $g = (g_1, g_2)^T$. Suppose the conditions in (3.2) are satisfied. In this case, $n_0 = 1$, and T_1 is the 6×6 matrix given by

$$T_1 = -\begin{bmatrix} 0 & 0 & & & & \\ 0 & 1 & & & & \\ D_1 h_{11}(0) & D_1 h_{12}(0) & -1 & 0 & & \\ D_1 h_{21}(0) & D_1 h_{22}(0) & 0 & 0 & & \\ D_2 h_{11}(0) & D_2 h_{12}(0) & 0 & 0 & -1 & 0 \\ D_2 h_{21}(0) & D_2 h_{22}(0) & 0 & 0 & 0 & 0 \end{bmatrix}.$$

By Theorem 2.4, (1.1) has a compactly supported distributional solution if and only if the linear equation

$$T_1 v = [\hat{g}_1(0), \hat{g}_2(0), 2D_1\hat{g}_1(0), 2D_1\hat{g}_2(0), 2D_2\hat{g}_1(0), 2D_2\hat{g}_2(0)]^T$$

has a solution v in $\mathbb{C}^6$. Let S be the set of all compactly supported distributional solutions of (1.1). There are two possible cases.

Case 1. Suppose $D_1h_{21}(0) = 0$ and $D_2h_{21}(0) = 0$. By Theorem 2.4, (1.1) has a compactly supported distributional solution if and only if

$$\hat{g}_1(0) = 0, \quad 2D_1\hat{g}_2(0) = \hat{g}_2(0)D_1h_{22}(0), \quad \text{and} \quad 2D_2\hat{g}_2(0) = \hat{g}_2(0)D_2h_{22}(0).$$

In this case, Corollary 2.6 confirms that $\dim(S) = 3$, since the dimension of $\ker(T_1)$ is 3.

Case 2. Suppose $D_1h_{21}(0) \neq 0$ or $D_2h_{21}(0) \neq 0$. In this case (1.1) has a compactly supported distributional solution if and only if $\hat{g}_1(0) = 0$ and

$$D_1h_{21}(0)\big(2D_2\hat{g}_2(0) - \hat{g}_2(0)D_2h_{22}(0)\big) = D_2h_{21}(0)\big(2D_1\hat{g}_2(0) - \hat{g}_2(0)D_1h_{22}(0)\big).$$

In this case, $\dim(S) = 2$ by the fact that the dimension of $\ker(T_1)$ is 2.

REFERENCES

[1] C. Cabrelli, C. Heil, and U. Molter, *Accuracy of lattice translates of several multidimensional refinable functions*, J. Approx. Theory, 95 (1998), pp. 5–52.

[2] J. Barros-Neto, *An Introduction to the Theory of Distributions*, Marcel Dekker, New York, 1973.

[3] A. S. Cavaretta, W. Dahmen, and C. A. Micchelli, *Stationary Subdivision*, Mem. Amer. Math. Soc. 93, Amer. Math. Soc., Providence, RI, 1991.

[4] C. K. Chui and X. Shi, *Continuous two-scale equations and dyadic wavelets,* Adv. Comput. Math., 2 (1994), pp. 185–213.

[5] A. Cohen, I. Daubechies, and G. Plonka, *Regularity of refinable function vectors*, J. Fourier Anal. Appl., 3 (1997), pp. 295–324.

[6] W. Dahmen and C. A. Micchelli, *Continuous refinement equations and subdivision,* Adv. Comput. Math., 1 (1993), pp. 1–37.

[7] I. Daubechies and J. C. Lagarias, *Two-scale difference equations:* I. *Existence and global regularity of solutions*, SIAM J. Math. Anal., 22 (1991), pp. 1388–1410.

[8] G. Derfel, N. Dyn, and D. Levin, *Generalized refinement equations and subdivision processes*, J. Approx. Theory, 80 (1995), pp. 272–297.

[9] T. B. Dinsenbacher and D. P. Hardin, *Nonhomogeneous refinement equations*, in Wavelets, Multiwavelets, and their Applications, A. Aldroubi and E. Lin, eds., AMS, Providence, RI, 1998, pp. 117–127.

[10] T. B. Dinsenbacher and D. P. Hardin, *Multivariate nonhomogeneous refinement equations*, J. Fourier Anal. Appl., 5 (1999), pp. 589–597.

[11] G. C. Donovan, J. S. Geronimo, D. P. Hardin, and P. R. Massopust, *Construction of orthogonal wavelets using fractal interpolation functions*, SIAM J. Math. Anal., 27 (1996), pp. 1158–1192.

[12] N. Dyn and A. Ron, *Multiresolution analysis by infinitely differentiable compactly supported functions,* Appl. Comput. Harmon. Anal., 2 (1995), pp. 15–20.

[13] B. Han and R. Q. Jia, *Multivariate refinement equations and convergence of subdivision schemes*, SIAM J. Math. Anal., 29 (1998), pp. 1177–1199.

[14] C. Heil and D. Colella, *Matrix refinement equations: Existence and uniqueness*, J. Fourier Anal. Appl., 2 (1996), pp. 75–94.

[15] R. Q. Jia, S. L. Lee, and A. Sharma, *Spectral properties of continuous refinement operators*, Proc. Amer. Math. Soc., 126 (1998), pp. 729–737.

[16] Q. T. Jiang and S. L. Lee, *Spectral properties of matrix continuous refinement operators*, Adv. Comput. Math., 7 (1997), pp. 383–399.

[17] Q. T. Jiang and Z. W. Shen, *On existence and weak stability of matrix refinable functions*, Constr. Approx., 15 (1999), pp. 337–353.

[18] K. Kabaya and M. Iri, *Sum of uniformly distributed random variables and a family of nonanalytic C^∞-functions,* Japan J. Appl. Math., 4 (1987), pp. 1–22.

[19] K. Kabaya and M. Iri, *On operators defining a family of nonanalytic C^∞-functions*, Japan J. Appl. Math., 5 (1988), pp. 333–365.

[20] M. H. Protter and C. B. Morrey, *A First Course in Real Analysis*, 2nd ed., Springer-Verlag, New York, 1977.

[21] V. A. Rvachev, *Compactly supported solutions of functional-differential equations and their applications,* Russian Math. Surveys, 45 (1990), pp. 87–120 .
[22] G. Strang and D. X. Zhou, *Inhomogeneous refinement equations*, J. Fourier Anal. Appl., 4 (1998), pp. 733–747.
[23] Q. Y. Sun, *Nonhomogeneous Refinement Equations: Existence, Regularity and Biorthogonality*, preprint, 1998.
[24] D. X. Zhou, *Existence of multiple refinable distributions*, Michigan Math. J., 44 (1997), pp. 317–329.

SIAM J. MATH. ANAL.
Vol. 32, No. 2, pp. 435–463

STUDY OF A CLASS OF REGULARIZATIONS OF $1/|X|$ USING GAUSSIAN INTEGRALS*

MARY BETH RUSKAI† AND ELISABETH WERNER‡

Abstract. This paper presents a comprehensive study of the functions

$$V_m^p(x) = \frac{pe^{x^p}}{\Gamma(m+1)} \int_x^\infty (t^p - x^p)^m e^{-t^p} dt$$

for $x > 0$, $m > -1$, and $p > 0$. For large x these functions approximate x^{1-p}. The case $p = 2$ is of particular importance because the functions $V_m^2(x) \approx 1/x$ can be regarded as one-dimensional regularizations of the Coulomb potential $1/|x|$ which are finite at the origin for $m > -\frac{1}{2}$.

The limiting behavior and monotonicity properties of these functions are discussed in terms of their dependence on m and p as well as x. Several classes of inequalities, some of which provide tight bounds, are established. Some differential equations and recursion relations satisfied by these functions are given. The recursion relations give rise to two classes of polynomials, one of which is related to confluent hypergeometric functions. Finally, it is shown that, for integer m, the function $1/V_m^2(x)$ is convex in x and this implies an analogue of the triangle inequality. Some comments are made about the range of p and m to which this convexity result can be extended and several related questions are raised.

Key words. Gaussian integrals, Coulomb potential, inequalities, Appell polynomials, confluent hypergeometric functions

AMS subject classifications. Primary, 33E20, 26D07; Secondary, 81V45

PII. S0036141099353758

1. Introduction.

1.1. Definitions and background.

In this paper we study the functions

$$V_m(x) = \frac{2e^{x^2}}{\Gamma(m+1)} \int_x^\infty (t^2 - x^2)^m e^{-t^2} dt, \quad m > -1, \tag{1}$$
$$V_{-1}(x) = \frac{1}{|x|}$$

and their generalizations

$$V_m^p(x) = \frac{pe^{x^p}}{\Gamma(m+1)} \int_x^\infty (t^p - x^p)^m e^{-t^p} dt, \tag{2}$$
$$V_{-1}^p(x) = x^{1-p}$$

for $0 < p < \infty$. These functions are well defined for $x > 0$ and can be extended to complex m with $\Re(m) > -1$. For $\Re(m) > -\frac{1}{2}$ they are also well defined for $x = 0$. Using symmetry or the equivalent forms (4) and (7) below, they can be extended to

*Received by the editors March 22, 1999; accepted for publication (in revised form) December 8, 1999; published electronically August 31, 2000.
http://www.siam.org/journals/sima/32-2/35375.html

†Department of Mathematics, University of Massachusetts Lowell, Lowell, MA 01854 (bruskai@cs.uml.edu). The work of this author was supported by National Science Foundation grant DMS-97-06981.

‡Department of Mathematics, Case Western Reserve University, Cleveland, OH 44106 (emw2@po.cwru.edu).

even functions on $\mathbf{R}$ or $\mathbf{R} \setminus \{0\}$. However, it suffices to consider only nonnegative x and in this paper we restrict ourselves to that. We also restrict ourselves to real m.

Letting $p = 2$ in (2) yields (1). However, because this case is more important in applications we often drop the superscript and simply write $V_m(x)$ for $V_m^2(x)$.

Our interest was motivated by studies of atoms in magnetic fields where these functions arise naturally for integer m. V_m can be regarded as a (two-dimensional) expectation of the (three-dimensional) Coulomb potential $1/|\mathbf{r}|$ with the state $\gamma_m(r,\theta) = \frac{1}{\sqrt{\pi m!}} e^{-im\theta} r^m e^{-r^2/2}$ (where we have used cylindrical coordinates $\mathbf{r} = (x, r, \theta)$ with the nonstandard convention $r = \sqrt{y^2+z^2}$ if $\mathbf{r} = (x, y, z)$ in rectangular coordinates). The state γ_m describes an electron in the lowest of the so-called "Landau levels" with angular momentum m in the direction of the field. In this context, it is natural to rewrite (1) in the form

$$V_m(x) = \frac{2}{\Gamma(m+1)} \int_0^\infty \frac{r^{2m} e^{-r^2}}{\sqrt{x^2+r^2}} r dr, \tag{3}$$

$$= \frac{1}{\Gamma(m+1)} \int_0^\infty \frac{u^m e^{-u}}{\sqrt{x^2+u}} du \tag{4}$$

for $m > -1$. In this form, it is easy to see that $V_m(x) \approx 1/|x|$ for large x. The importance of V_m goes back at least to Schiff and Snyder [20] and played an essential role in the Avron, Herbst, and Simon [2] study of the energy asymptotics of hydrogen in a strong magnetic field. More recently, work in astrophysics and the work of Lieb, Solovej, and Yngvason [13, 14] on asymptotics of many-electron atoms in strong magnetic fields has renewed interest in this subject. Motivated by the work in [13, 14], Brummelhuis and Ruskai [7, 8] have developed one-dimensional models of many-electron atoms in strong magnetic fields using the functions $V_m(x)$ as one-dimensional analogues of the Coulomb potential.

In the case of many-electron atoms, the antisymmetry required by the Pauli exclusion principle suggests replacing the simple "one-electron" expectation above by an N-electron analogue in which the state γ_m is replaced by a Slater determinant of such states. This is discussed in detail in [8] where it is shown that, in the simple case corresponding to $m = 0 \ldots N-1$, the analogous one-dimensional potentials have the form

$$V_{\mathrm{av}}^N(x) = \frac{1}{N} \sum_{m=0}^{N-1} V_m(x). \tag{5}$$

In section 3 we obtain recursion relations for V_m which, in addition to being of considerable interest in their own right, are extremely useful for studying potentials of the form (5).

For $m = 0$ the function

$$\frac{1}{\sqrt{2}} V_0\left(\frac{x}{\sqrt{2}}\right) = e^{x^2/2} \int_x^\infty e^{-t^2/2} dt \tag{6}$$

occurs in many other contexts and is sometimes called the "Mills ratio" [15]. Although it has been extensively studied, the class of inequalities we consider in section 4.1 appears to be new (although some of our bounds coincide with known inequalities in other classes) and the realization that $1/V_0(x)$ is convex seems to be relatively recent [18, 19, 7].

The replacement of x^2 by x^p in (6) has been considered by Gautschi [10] and Mascioni [12], who (after seeing the preprint [16]) extended the results of section 4.1 to this situation.

For the analysis of this generalization, it is useful to observe that (2) can be rewritten as

$$V_m^p(x) = \frac{1}{\Gamma(m+1)} \int_0^\infty \frac{u^m e^{-u}}{(x^p+u)^{\frac{p-1}{p}}} du. \tag{7}$$

In this form, it is easy to verify that $V_m^p(x) \approx x^{p-1}$ for large x and that for $p = 1$, $V_m^1(x) = 1$ for all m.

Our first result shows that $V_m^p(x)$ is continuous in m and that our definition for $m = -1$ is natural.

PROPOSITION 1. *For all $x > 0$, $\lim_{m\to -1^+} x^{p-1} V_m^p(x) = 1$.*

Proof. Note that

$$\frac{x^{p-1}}{(x^p+u)^{\frac{p-1}{p}}} = \frac{1}{\left(1+\frac{u}{x^p}\right)^{\frac{p-1}{p}}}$$

so that (7) implies

$$1 - x^{p-1} V_m^p(x) = \frac{1}{\Gamma(m+1)} \int_0^\infty u^m e^{-u} \left[1 - \left(1+\frac{u}{x^p}\right)^{\frac{1-p}{p}}\right] du. \tag{8}$$

Since $\Gamma(m+1)$ becomes infinite as $m \to -1$, the desired result follows if the integral on the right-hand side above remains finite. To see that this is true, it is convenient to let $g(z) = \frac{1}{z}(1-(1+z)^{\frac{1-p}{p}})$ and note that (8) implies

$$\left|1 - x^{p-1} V_m^p(x)\right| \le \frac{1}{\Gamma(m+1)} \int_0^\infty \frac{u^{m+1} e^{-u}}{x^p} \left|g\left(\frac{u}{x^p}\right)\right| du. \tag{9}$$

It is easy to see that the large u portion of this integral causes no problems since $|g(z)|$ is bounded by a polynomial in z when $z > 1$. (For $p \ge 1$, it is bounded uniformly by 1; for $0 < p < 1$, it is bounded by a polynomial, namely, $|g(z)| \le 1 + (1+z)^k$, where $k \in \mathbf{N}$, $k \ge \frac{1}{p}$.) To see that it is also well behaved for small u, we first note that for $p > 0$, $(1+z)^{\frac{1-p}{p}}$ is analytic for $\Re(z) > -1$. Then $g(z)$ has a removable singularity at $z = 0$ and can be extended to an analytic function on $\Re(z) > -1$. Thus, for small u, the integrand behaves like $x^{-p} u^{m+1} e^{-u}$, which ensures that the integral in (9) is finite for $m = -1$. □

The rest of this paper is organized as follows. In the next part of this section we summarize the properties of V_m in the important case $p = 2$. We then conclude the introduction with a summary of convexity results, including some open questions. In section 2 we state and prove the basic properties of V_m^p for general p. In section 3 we derive recursion relations for V_m^p and study their consequences. Among these is a connection with confluent hypergeometric functions. In section 4.1 we prove some optimal bounds for V_0. The optimal upper bound had been established earlier independently by Wirth [19] and by Szarek and Werner [18], who also showed that the upper bound is equivalent to the convexity of $1/V_0$. In section 4.2 we discuss several classes of inequalities, beginning with optimal bounds on $V_0(x)$. We then consider optimal bounds on the ratio $R_m(x) = V_m(x)/V_{m-1}(x)$ and show that these have

important consequences. In particular, we show that the upper bound is equivalent to the convexity (in x) of $1/V_m(x)$ and that the ratios increase with x. Proofs of the ratio bounds are then given in section 5 where we also consider extensions to other p. Because the proof of the ratio bounds is via induction on m, the results of sections 4 and 5 are established only for integer m. However, we believe that they hold for all $m > -1$.

1.2. Properties of $V_m(x)$. We now summarize some properties of $V_m(x)$ along with comments about the history and brief remarks about the proofs. Unless otherwise stated, these properties hold for $m > -1$ and $x > 0$.

(a) $$\frac{1}{\sqrt{x^2+m}} > V_m(x) > \frac{1}{\sqrt{x^2+m+1}},$$

where the first inequality holds for $m > 0$ and the second for $m > -1$.
To prove the upper bound, which appears to be new, observe that $\mu = [u^{m-1}e^{-u}/\Gamma(m)]du$ is a probability measure on $(0,\infty)$. For fixed x, one can then apply Jensen's inequality to the concave function $f_x(u) = u(u+x^2)^{-1/2}$ to obtain

$$V_m(x) = \frac{1}{m}\int_0^\infty f_x(u)d\mu(u) \le \frac{1}{m} f_x\left(\int_0^\infty \frac{u^m e^{-u}du}{\Gamma(m)}\right)$$

$$= \frac{1}{m} f_x(m) = \frac{1}{\sqrt{x^2+m}}.$$

The lower bound was proved earlier (at least for integer m) by Avron, Herbst, and Simon [2], who applied a similar argument to the probability measure $[u^m e^{-u}/\Gamma(m+1)]du$ and the convex function $f_x(u) = (u+x^2)^{-1/2}$.

(b) $V_m(x)$ is decreasing in m. In particular, $V_{m+1}(x) < V_m(x) < \frac{1}{x}$.
The first inequality follows easily from property (a), which implies $V_m(x) < \frac{1}{\sqrt{x^2+m}} < V_{m-1}(x)$. Alternatively, one could use integration by parts on (4). The second inequality is easily verified from the integral representation (4). That $V_m(x)$ also decreases with m for noninteger jumps is more difficult, and the proof is postponed to section 2, where it follows from the more general Theorem 6.

(c) The expression $mV_m(x)$ is increasing in $m > -1$, $m \in \mathbf{R}$.
For integer jumps this holds for $m \ge -1$. Indeed, it is obvious that $-V_{-1} < 0 \cdot V_0 < V_1$. For integer jumps with $m \ge 1$, one can use property (a) to see that

$$mV_m(x) > \frac{m}{\sqrt{x^2+m+1}} > \frac{m-1}{\sqrt{x^2+m-1}} > (m-1)V_{m-1}(x). \tag{10}$$

The proof for general m is postponed to Theorem 6 in section 2. The fact that $V_m(x)$ is decreasing in m, while $mV_m(x)$ is increasing gives an indication of the delicate behavior of V_m.

(d) For $m > -1/2$, the definition of $V_m(x)$ can be extended to $x = 0$ and

$$V_m(0) = \frac{\Gamma(m+\frac{1}{2})}{\Gamma(m+1)}. \tag{11}$$

For integer m, this becomes

$$V_m(0) = \frac{(2m)!}{2^{2m}(m!)^2}\sqrt{\pi} = \frac{1\cdot 3\cdot 5\ldots(2m-1)}{2\cdot 4\cdot 6\ldots(2m)}\sqrt{\pi}, \tag{12}$$

while for large m Stirling's formula implies

$$V_m(0) \approx \left(\frac{m-\frac{1}{2}}{m}\right)^m \left(\frac{e}{m}\right)^{1/2} \approx \frac{1}{\sqrt{m}}, \tag{13}$$

which is consistent with property (a). Boyd [6, 15] has proved the more precise estimates

$$\frac{\sqrt{m+\frac{3}{4}+\frac{1}{32m+48}}}{m+\frac{1}{2}} < V_m(0) < \frac{1}{\sqrt{m+\frac{1}{4}+\frac{1}{32m+32}}}. \tag{14}$$

(e) For all $m \geq 0$, V_m satisfies the differential equation

$$V_m'(x) = 2x\left(V_m - V_{m-1}\right). \tag{15}$$

This can easily be verified using integration by parts in (4).

(f) For each fixed $m \geq 0$, $V_m(x)$ is decreasing in x.
This follows directly from (b) and (e).

(g) For $a > 0$, the expression $aV_m(ax)$ increases with a. Hence $aV_m(ax) > V(x)$ when $a > 1$ and $aV_m(ax) < V(x)$ when $a < 1$.
This property follows easily from the definition (3) or (4) and the observation that $\frac{a}{\sqrt{a^2x^2+u}} = \frac{1}{\sqrt{x^2+\frac{u}{a^2}}}$ is increasing in a. It is used in the proof of Theorem 6 and is important in the study of one-dimensional models for atoms in magnetic fields in which the electron-electron interaction takes the form of convex combinations of $\frac{1}{\sqrt{2}}V_m(\frac{|x_j-x_k|}{\sqrt{2}})$.

(h) $V_0(x)$ is convex in $x > 0$; however, $V_m(x)$ is *not* convex when $m > \frac{1}{2}$.
For $m = 0$, the differential equation (15) becomes $V_0'(x) = 2[xV_0 - 1]$. Since $xV_0 = \int_0^\infty \frac{e^{-u}}{\sqrt{1+u/x^2}}du$ is increasing for $x > 0$, it follows that $V_0(x)$ is convex.
When $m > \frac{1}{2}$, it follows from (15) and (b) that $\lim_{x\to 0} V_m'(x) = 0$. Since V_m' is negative, V_m' must decrease on some small interval $(0, x_0)$. One can also show $\lim_{x\to\infty} V_m'(x) = 0$, so that one expects that there is an x_1 such that V_m is concave on $(0, x_1)$ and convex on (x_1, ∞). In section 3 we will see that the convexity is recovered for the averaged potential V_m^{av}.

(i) For integer m, $1/V_m(x)$ is convex in $x > 0$.
This will be proved in section 4.3 as Theorem 23. For large x, $1/V_m(x) \approx x$ so that the deviation from linearity is very small and the second derivative close to zero. This makes the proof quite delicate and lengthy.
The convexity of $1/V_m(x)$ can be rewritten as

$$\frac{1}{\frac{1}{2}V_m\left(\frac{x+y}{2}\right)} \leq \frac{1}{V_m(x)} + \frac{1}{V_m(y)}.$$

Using property (g) with $a = \frac{1}{2}$, one easily finds that the convexity of $1/V_m(x)$ implies

$$\frac{1}{V_m(x+y)} \leq \frac{1}{V_m(x)} + \frac{1}{V_m(y)}.$$

This subadditivity inequality plays the role of the triangle inequality in applications. (See, e.g., [7].)

(j) Asymptotic estimates.
For large x, it follows from property (a) that

$$\frac{m}{2(x^2+m)^{3/2}} \le \frac{1}{x} - V_m(x) < \frac{m+1}{2x^3}. \tag{16}$$

The asymptotic expansion

$$V_m(x) = \frac{1}{x} - \frac{m+1}{2x^3} + \frac{3(m+2)(m+1)}{8x^5} + O\left(\frac{1}{x^7}\right) \tag{17}$$

can easily be obtained from (4). For details, let $p = 2$ in the proof of Proposition 7, which gives a similar expansion for $p > 1$.
It follows from properties (a) and (c) that $V_m(x)$ decreases monotonically to zero for each fixed x as $m \to \infty$. In fact, since $V_m(x)$ is decreasing in x for all m, it suffices to show this for $x = 0$, which is easy since $V_m(0) < m^{-1/2}$.

(k) The Fourier transform is given by

$$\widehat{V}_m(\xi) \equiv \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} V_m(x)\, e^{-ix\xi} dx = \frac{4^{m+1}}{\sqrt{2\pi}} \int_0^{\infty} \frac{s^m\, e^{-s}}{(|\xi|^2 + 4s)^{m+1}}\, ds. \tag{18}$$

This follows from (3) and the standard formula (e.g., see (v) on page 131 of [17]) $\mathcal{F}(\frac{1}{\sqrt{|x|^2+|w|^2}})(\xi) = \frac{1}{\sqrt{2\pi}} \int_0^{\infty} \frac{1}{s} e^{-\frac{1}{2}(s+|w|^2|\xi|^2/s)}\, ds$ after a change in the order of integration.

1.3. Convexity summary. For large x, $V_m^p(x) \approx x^{1-p}$, which is convex in x for $p > 1$ and concave for $p < 1$. For $m > -\frac{1}{2}$ these convexity properties cannot be extended to $V_m^p(x)$ on all of $(0, \infty)$; they would be inconsistent with the differential equation and monotonicity properties in Proposition 5. However, as discussed after Proposition 11, the averaged potentials $V_{\mathrm{av}}^{p,N}$ have the same convexity as x^{1-p} on the half-line.

The convexity of $1/V_m^p(x)$ is the motivation for sections 4 and 5. This question is already delicate for $p = 2$ and its verification becomes increasingly difficult for larger p. Although, as discussed in section 5.4, we have evidence that convexity holds for all $p \ge 2$, our methods give this result only for a limited range of p. Moreover, because our proof is inductive, we have established convexity and ratio bounds of section 4.2 only for integer m. It would be interesting to find another approach which would extend these results to noninteger m and all $p \ge 2$. Since $1/V_m^p(x) \approx x^{p-1}$, which is concave for $1 < p < 2$, we cannot expect convexity of $1/V_m^p$ in this range.

As discussed above, one important consequence of the convexity of $1/V_m(x)$ is an analogue of the triangle inequality. For all values of p we have $[V_m^p(x)]^{\frac{1}{1-p}} \approx x$ for large x, which suggests a triangle inequality of the form

$$[V_m^p(x+y)]^{\frac{1}{1-p}} \le [V_m^p(x)]^{\frac{1}{1-p}} + [V_m^p(y)]^{\frac{1}{1-p}}.$$

It would be interesting to know the range of p (and m) for which this holds. For $p > 2$ the convexity of $1/V_m^p(x)$ implies only the weaker inequality

$$[V_m^p(x+y)]^{\frac{1}{1-p}} \le 2^{\frac{p-2}{p-1}} \left([V_m^p(x)]^{\frac{1}{1-p}} + [V_m^p(y)]^{\frac{1}{1-p}} \right).$$

Finally, one could also ask if $V_m(x)$ is convex in m. In particular, is $2V_m(x) \le V_{m+1}(x) + V_{m-1}(x)$ or, equivalently by (15), is $V_m'(x)$ increasing in m?

2. General p. We now study the basic properties of V_m^p in detail. As one would expect from $V_m^p(x) \approx x^{1-p}$, the behavior of V_m^p is often quite different for $p > 1$ and $p < 1$. At the boundary, $p = 1$, $V_m^1(x) = 1$ for all x. Proposition 2 describes the monotonicity and limiting behavior of $V_m^p(x)$ as p varies with m and x fixed. Proposition 3 gives a simple expression for V_m^p in the special case that $1/p$ is an integer.

The next four results generalize properties of section 1.2 to general p. Proposition 4 generalizes the inequalities from property (a); Proposition 5 generalizes properties (d), (e), (f), and (g); and Theorem 6 extends the monotonicity properties (b) and (c). Moreover, the proof of monotonicity for noninteger jumps is provided here. Finally, Proposition 7 gives the asymptotic behavior of $V_m^p(x)$ for large x when $p > 1$.

PROPOSITION 2. *Let $m > -1$ and $x > 0$ be fixed. Then we have*

(i) $\lim_{p\to 0} V_m^p(x) = \infty$;

(ii) *for all $x \geq 1$, V_m^p is decreasing in p; moreover,*

$$\text{if } x > 1, \quad \lim_{p\to\infty} V_m^p(x) = 0, \quad \text{and}$$

$$\text{if } x = 1, \quad \lim_{p\to\infty} V_m^p(1) = \frac{1}{\Gamma(m+1)} \int_0^\infty \frac{u^m e^{-u} du}{1+u};$$

(iii) *for all $0 < x < 1$ and $m > 0$, $\lim_{p\to\infty} V_m^p(x) = \frac{1}{m}$.*

Proof. We use the expression (7) for V_m^p.

(i) Since $\lim_{p\to 0}(x^p + u)^{1/p} = \infty$ for $x > 1$,

$$\lim_{p\to 0} V_m^p(x) \geq \lim_{p\to 0} \frac{1}{\Gamma(m+1)} \int_1^3 \frac{u^m e^{-u}(x^p+u)^{\frac{1}{p}} du}{x^p + u} = \infty.$$

(ii) Differentiating (7) yields

$$\frac{d}{dp} V_m^p(x) = \frac{1}{\Gamma(m+1)} \int_0^\infty \frac{u^m e^{-u}}{p^2 (x^p+u)^{\frac{2p-1}{p}}} \times \Big[(1-p) x^p \ln(x^p) - (x^p+u)\ln(x^p+u) \Big] du. \tag{19}$$

For $x = 1$ or $p = 1$, the first term in square brackets above is zero, leaving a quantity which is clearly negative. When both $x > 1$ and $p > 1$, both terms in (19) are clearly negative. When $x > 1$ and $0 < p < 1$, the quantity in square brackets in (19) is negative since

$$\begin{aligned}(1-p)x^p \ln(x^p) - (x^p+u)\ln(x^p+u) \\ \leq x^p \ln(x^p) - (x^p+u)\ln(x^p+u) < 0.\end{aligned}$$

The last inequality follows from the fact that the function $f(w) = w \ln w$ is increasing for $w > 1$. Thus, $\frac{d}{dp} V_m^p(x) \leq 0$ for all $x \geq 1$ and for all $p \in (0, \infty)$.

(iii) Since $\lim_{p\to\infty}(x^p + u)^{1/p} = 1$ for $x < 1$,

$$\lim_{p\to\infty} V_m^p(x) = \frac{1}{\Gamma(m+1)} \int_0^\infty u^{m-1} e^{-u} du = \frac{1}{m}. \qquad \square$$

Remark. The behavior for $x < 1$ depends upon m. For "small" m, there is a p_0 such that V_m^p decreases (below 1) on $(0, p_0)$ and increases on (p_0, ∞) to $\frac{1}{m}$. For "big" m, V_m^p simply decreases to $\frac{1}{m}$ as p increases.

The next result shows that in the special case that $1/p$ is an integer, V_m^p reduces to a polynomial in $x^p = x^{1/n}$ of degree $n-1$.

PROPOSITION 3. *For $n \in \mathbf{N}$, $n \geq 2$,*

$$V_m^{1/n}(x) = \frac{1}{\Gamma(m+1)} \sum_{k=0}^{n-1} \binom{n-1}{k} \Gamma(m+n-k)\ x^{k/n}$$

for all $x \geq 0$ and $m > -1$.

Proof. It follows from (7) that for $p = \frac{1}{n}$

$$V_m^{1/n}(x) = \frac{1}{\Gamma(m+1)} \int_0^\infty \left(x^{1/n} + u\right)^{n-1} u^m e^{-u} du.$$

When n is an integer ≥ 2, the result then follows easily from the binomial expansion applied to $(x^{\frac{1}{n}} + u)^{n-1}$ and the definition of the Γ-function.

PROPOSITION 4. *For all $x > 0$*

$$\frac{1}{(x^p + m + 1)^{\frac{p-1}{p}}} \leq V_m^p(x) \leq \frac{1}{(x^p + m)^{\frac{p-1}{p}}} \quad \text{for } p > 1,$$

where the first inequality holds for $m > -1$ and the second for $m \geq 0$.

$$(x^p + m + 1)^{\frac{1-p}{p}} \geq V_m^p(x) \geq (x^p + m)^{\frac{1-p}{p}} \quad \text{for } \frac{1}{2} \leq p < 1,$$

where the first inequality holds for $m > -1$ and the second for $m \geq 0$.

$$V_m^p(x) \geq (x^p + m + 1)^{\frac{1-p}{p}} \quad \text{for } 0 < p \leq \frac{1}{2},$$

and the inequality holds for $m > -1$.

Proof. The proofs are done using Jensen's inequality as in property (a) of section 1.2.

PROPOSITION 5. *For all $x > 0$*

(i) *for $m > -\frac{1}{p}$, $V_m^p(0)$ is defined and $V_m^p(0) = \frac{\Gamma(m+\frac{1}{p})}{\Gamma(m+1)}$;*

(ii) *for all $m \geq 0$, $x > 0$, V_m^p satisfies the differential equation*

$$\frac{d}{dx} V_m^p(x) = p x^{p-1} \left(V_m^p(x) - V_{m-1}^p(x)\right); \tag{20}$$

(iii) *for all $m > -1$, V_m^p is decreasing in x if $p > 1$, identically equal to 1 for all x if $p = 1$, and increasing in x if $p < 1$;*

(iv) *let $m > -1$ and $x > 0$; for $a > 0$, the expression $a^{p-1} V_m^p(ax)$ increases in a if $p > 1$ and decreases in a if $p < 1$.*

Proof. The proofs are straightforward extensions of those given in section 1.2. In (iii), one can verify that V_m^p is also increasing for $0 < p < \frac{1}{2}$ by computing the derivative directly.

THEOREM 6. *For each fixed $x > 0$, and for m in the region $m > -1$,*

(i) *$V_m^p(x)$ is strictly decreasing in m for $p > 1$ and strictly increasing in m for $p < 1$;*

(ii) *$mV_m^p(x)$ is strictly increasing in m for $p > 1$ and strictly decreasing in m for $p < 1$.*

Proof. To prove (i) we differentiate (7) to get

$$\frac{d}{dm}V_m^p(x) = \frac{1}{\Gamma(m+1)}\int_0^\infty \frac{u^m \ln u\ e^{-u}}{(x^p+u)^{1-\frac{1}{p}}}du - V_m^p(x)\frac{\Gamma'(m+1)}{\Gamma(m+1)}. \tag{21}$$

Using the same procedure as that used (see, e.g., [1, 11]) to obtain the standard integral representation

$$\psi(z) \equiv \frac{\Gamma'(z)}{\Gamma(z)} = \int_0^\infty \left[\frac{e^{-s}}{s} - \frac{1}{s(1+s)^z}\right] ds, \tag{22}$$

one finds

$$\begin{aligned}
&\frac{1}{\Gamma(m+1)}\int_0^\infty \frac{u^m \ln u\ e^{-u}}{(x^p+u)^{1-\frac{1}{p}}}du \\
&= \frac{1}{\Gamma(m+1)}\int_{s=0}^\infty \frac{ds}{s}\int_0^\infty [e^{-s}-e^{-su}]\frac{e^{-u}u^m}{(x^p+u)^{1-\frac{1}{p}}}du \\
&= V_m^p(x)\int_0^\infty \frac{e^{-s}}{s}ds - \frac{1}{\Gamma(m+1)}\int_0^\infty \frac{ds}{s(s+1)^{m+\frac{1}{p}}}\int_0^\infty \frac{e^{-w}w^m dw}{[x^p(s+1)+w]^{1-\frac{1}{p}}} \\
&= V_m^p(x)\int_0^\infty \frac{e^{-s}}{s}ds - \int_0^\infty V_m^p(x(s+1)^{\frac{1}{p}})\frac{ds}{s(s+1)^{m+\frac{1}{p}}},
\end{aligned}$$

where we made the change of variable $w = (s+1)u$ to obtain $V_m^p(x(s+1)^{\frac{1}{p}})$. Now we use Proposition 5 with $a = (s+1)^{\frac{1}{p}} > 1$ to obtain

$$\begin{aligned}
\frac{1}{\Gamma(m+1)}\int_0^\infty \frac{u^m \ln u\ e^{-u}}{(x^p+u)^{1-\frac{1}{p}}}du &\le \int_0^\infty V_m^p(x)\left(\frac{e^{-s}}{s} - \frac{1}{s(s+1)^{m+1}}\right)ds \\
&= V_m^p(x)\ \psi(m+1)
\end{aligned}$$

when $p > 1$. For $p < 1$, Proposition 5 gives the inequality in the opposite direction. Hence inserting the result in (21) yields

$$\frac{d}{dm}V_m(x)\begin{cases} <0 & \text{if } p>1, \\ >0 & \text{if } p>1. \end{cases}$$

To prove (ii) it is slightly more convenient to consider the logarithmic derivative $\frac{d}{dm}\ln[mV_m^p(x)]$ and show that it is positive for $p > 1$ and negative for $p < 1$. Proceeding as above, we find for $p > 1$

$$\begin{aligned}
\frac{d}{dm}\ln[mV_m^p(x)] &= \frac{1}{m} + \frac{\frac{d}{dm}[V_m^p(x)]}{V_m^p(x)} \\
&= \frac{1}{m} + \int_0^\infty \frac{e^{-s}}{s}ds - \frac{1}{V_m^p(x)}\int_0^\infty \frac{V_m^p\left[x(s+1)^{1/p}\right]}{s\ (s+1)^{m+\frac{1}{p}}}ds - \psi(m+1) \\
&< \frac{1}{m} + \int_0^\infty \frac{e^{-s}}{s}ds - \int_0^\infty \frac{1}{s(s+1)^m}ds - \psi(m+1) \\
&= \frac{1}{m} + \psi(m) - \psi(m+1) \\
&= \frac{1}{m} - \int_0^\infty \frac{1}{(s+1)^{m+1}} = 0,
\end{aligned}$$

where we have used (22) and the following inequality with $a = (s+1)^{1/p}$.

$$V_m^p(ax) \left\{ \begin{array}{c} < \\ > \end{array} \right\} a V_m^p(x) \quad \text{for} \quad \left\{ \begin{array}{c} p > 1 \\ p < 1 \end{array} \right\} \tag{23}$$

for all $a \geq 1$. This is easily verified and implies that the inequality proved above for $\frac{d}{dm} \ln [m V_m^p(x)]$ is reversed when $p < 1$. □

The following result gives the asymptotic behavior of $V_m^p(x)$ for large x.

PROPOSITION 7. *For $p > 1$, $V_m^p(x)$ has the asymptotic expansion*

$$\frac{1}{x^{p-1}} - \frac{(p-1)(m+1)}{p\ x^{2p-1}} + \frac{(2p^2-3p+1)(m^2+3m+2)}{2p^2\ x^{3p-1}} + O\left(\frac{1}{x^{4p-1}}\right).$$

Proof. This follows from (7) since

$$\begin{aligned} V_m^p(x) &= \frac{1}{\Gamma(m+1)\ x^{p-1}} \int_0^\infty \frac{u^m e^{-u}}{\left(1+\frac{u}{x^p}\right)^{\frac{p-1}{p}}} du \\ &= \frac{1}{\Gamma(m+1)\ x^{p-1}} \int_0^\infty u^m e^{-u} \left[1 - \frac{(p-1)\ u}{p\ x^p} + \frac{(p-1)(2p-1)\ u^2}{2p^2\ x^{2p}} + \cdots\right] du \\ &= \frac{1}{x^{p-1}} \left[1 - \frac{(p-1)\ \Gamma(m+2)}{p\ \Gamma(m+1)\ x^p} + \frac{(2p^2-3p+1)\ \Gamma(m+3)}{(2p^2)\ \Gamma(m+1)\ x^{2p}} + O\left(\frac{1}{x^{3p}}\right)\right]. \end{aligned}$$

3. Recursion relations and their consequences.

3.1. Recursion relations for V_m^p. Although the case $p = 2$ is of primary interest in applications, we continue to study general p is this section, as the proofs for general p are identical to those for $p = 2$. In these recursions, our convention that $V_{-1}^p(x) = x^{1-p}$ plays an important role.

PROPOSITION 8. *For all $m \in \mathbf{R}$, $m \geq 1$, for all $x > 0$,*

$$V_m^p(x) = \frac{1}{m} \left[\left(m - 1 + \frac{1}{p} - x^p\right) V_{m-1}^p(x) + x^p V_{m-2}^p(x)\right]. \tag{24}$$

Proof. For $m = 1$, one gets that

$$V_1^p(x) = p e^{x^p} \left[\left(\frac{1}{p} - x^p\right) \int_x^\infty e^{-t^p} dt + x\right] = \left(\frac{1}{p} - x^p\right) V_0^p(x) + x^{p-1} V_{-1}^p(x).$$

For $m > 1$, using (2) and integration by parts, we find

$$\begin{aligned} V_m^p(x) &= \frac{p e^{x^p}}{\Gamma(m+1)} \int_x^\infty (t^p - x^p)^{m-1} (t^p - x^p) e^{-t^p} dt \\ &= \frac{p e^{x^p}}{m\Gamma(m)} \left[-x^p \int_x^\infty e^{-t^p} (t^p - x^p)^{m-1} dt \right. \\ &\qquad \left. + \frac{1}{p} \int_x^\infty e^{-t^p} ((t^p - x^p)^{m-1} + (m-1) p t^p (t^p - x^p)^{m-2}) dt\right] \\ &= \frac{p e^{x^p}}{m\Gamma(m)} \left[\left(m - 1 + \frac{1}{p} - x^p\right) \int_x^\infty e^{-t^p} (t^p - x^p)^{m-1} dt \right. \\ &\qquad \left. + \ (m-1) x^p \int_x^\infty (t^p - x^p)^{m-2} e^{-t^p} dt\right] \\ &= \frac{1}{m} \left[\left(m - 1 + \frac{1}{p} - x^p\right) V_{m-1}^p(x) + + x^p V_{m-2}^p(x)\right]. \end{aligned}$$

Repeated application of (24) gives a useful corollary. For $m \in \mathbf{R}$, let $\lfloor m \rfloor$ denote the "floor" of m, i.e., the largest natural number less than or equal to m.

COROLLARY 9. *Let $m \in \mathbf{R}$, $m \geq 1$, and let $n \in \mathbf{N}$ such that $n \leq \lfloor m \rfloor$. Then*

$$V_m^p(x) = \frac{1}{pm}\Big[(1-px^p)V_{m-1}^p(x) + V_{m-2}^p(x) + \cdots [p(m-n)+1]V_{m-n}^p(x) + px^pV_{m-n-1}^p(x)\Big]. \tag{25}$$

In particular, if m is a positive integer, then

$$V_m^p(x) = \frac{1}{pm}\left[(1-px^p)V_{m-1}^p(x) + \sum_{k=0}^{m-2} V_k^p(x) + px^pV_{-1}^p(x)\right]. \tag{26}$$

The expression (26) is well defined for $x = 0$. Putting $x = 0$ and using Proposition 5(i), we obtain the (presumably well-known) identity

$$\frac{\Gamma(m+\frac{1}{p})}{\Gamma(m+1)} = \frac{1}{pm}\sum_{k=0}^{m-1}\frac{\Gamma(k+\frac{1}{p})}{\Gamma(k+1)}. \tag{27}$$

3.2. Averaged potentials. These recursion relations are quite useful for studying the average of the first N of the V_m. For N a positive integer, we extend (5) to

$$V_{\text{av}}^{p,N}(x) = \frac{1}{N}\sum_{m=0}^{N-1} V_m^p(x). \tag{28}$$

Note that for $p = 1$, $V_{\text{av}}^{1,N}(x) = 1$ for all $x \geq 0$.

The next result follows immediately from (26).

COROLLARY 10. $V_{\text{av}}^{p,N}(x) = pV_N^p(x) - \frac{px^p}{N}\left[V_{-1}^p(x) - V_{N-1}^p(x)\right]$.

For the important case $p = 2$, this reduces to

$$V_{\text{av}}^N(x) = 2V_N(x) - \frac{2x^2}{N}\left[V_{-1}(x) - V_{N-1}(x)\right]. \tag{29}$$

The function $V_0(|x|)$ is convex on $(0,\infty)$ but has a cusp at $x = 0$. However, as discussed in property (f), for higher m both the convexity and cusp are lost. Thus, for higher m, the V_m are somewhat smoother than one might want for one-dimensional approximations to the Coulomb potential. The next result, although straightforward, is important because it implies that the averaged potentials $V_{\text{av}}^N(x)$ retain the cusp and convexity properties of V_0 near the origin.

PROPOSITION 11. *The function $V_{\text{av}}^N(x)$ is convex for all $x > 0$ and*

$$\lim_{x\to 0+}\frac{d}{dx}V_{\text{av}}^N(x) = -\frac{2}{N}.$$

Proof. Using (5) and (15) one finds

$$\frac{d}{dx}NV_{\text{av}}^N(x) = \sum_{m=0}^{N-1} 2x[V_m(x) - V_{m-1}] = 2x[V_{N-1} - V_{-1}] = 2xV_{N-1} - 2.$$

Therefore, to show that $V_{\rm av}^N$ is convex, we need to show that

$$xV_{N-1} = \frac{1}{\Gamma(N)} \int_0^\infty \frac{u^{N-1}e^{-u}du}{(1+\frac{u}{x^2})^{\frac{1}{2}}}$$

is increasing. This holds as, for $x > 0$, the function $h_u(x) = [1 + u/x^2]^{-1/2}$ is increasing.

Similarly, one can show that for $p > 1$, $V_{\rm av}^{p,N}$ is convex on $(0, \infty)$ and

$$\lim_{x\to 0+} \frac{d}{dx} V_{\rm av}^{p,N}(x) = -\frac{p}{N}.$$

For $p < 1$, the derivative becomes infinite at the origin; however, concavity of $V_{\rm av}^{p,N}$ on $(0, \infty)$ still holds.

3.3. Polynomials defined by recursion. We now observe that by repeatedly using (24) to eliminate the V_m^p with the largest value of m from (25) allows us to write V_m^p in terms of the two "lowest" functions (e.g., V_0^p and V_{-1}^p in the case of integer m) and that the coefficients in such expressions define two classes of polynomials related to confluent hypergeometric functions. We discuss the properties of these polynomials in some detail. First, we make the statement above explicit.

COROLLARY 12. *For $m \geq 1$ there are polynomials $P_m^p(y)$ and $Q_m^p(y)$ of degree $\lfloor m \rfloor$ such that*

$$V_m^p(x) = P_m^p(x^p)V_{m-\lfloor m\rfloor}^p(x) + x^p Q_{m-1}^p(x^p)V_{m-\lfloor m\rfloor-1}^p(x). \tag{30}$$

In the case of integer m (30) becomes

$$\begin{aligned} V_m^p(x) &= P_m^p(x^p)V_0^p(x) + x^p Q_{m-1}^p(x^p)V_{-1}^p(x) \\ &= P_m^p(x^p)V_0^p(x) + xQ_{m-1}^p(x^p), \end{aligned} \tag{31}$$

where the second expression follows from our convention $V_{-1}^p(x) = x^{1-p}$. We define $P_m^p(y) = 1$ for $m \in [0, 1)$ and $Q_m^p(y) = 0$ for $m \in [-1, 0)$. Then (30) holds trivially for $m \in [0, 1)$.

Proof. The desired polynomials are defined recursively. First let

$$P_m^p(y) = \frac{1}{m}\left(m - 1 + \frac{1}{p} - y\right) \quad \text{for} \quad m \in [1, 2) \quad \text{and} \tag{32}$$

$$Q_m^p(y) = \frac{1}{m+1} \quad \text{for} \quad m \in [0, 1). \tag{33}$$

Then (30) holds because it is equivalent to (24) for $m \in [1, 2)$. For $m \geq 2$ define $P_m^p(y)$ by

$$P_m^p(y) = \frac{1}{m}\left[\left(m - 1 + \frac{1}{p} - y\right) P_{m-1}^p(y) + y\ P_{m-2}^p(y)\right] \tag{34}$$

and for $m \geq 1$ define $Q_m^p(y)$ by

$$Q_m^p(y) = \frac{1}{m+1}\left[\left(m + \frac{1}{p} - y\right) Q_{m-1}^p(y) + y\ Q_{m-2}^p(y)\right]. \tag{35}$$

It is straightforward to use induction to check that (24) yields (30). ∎

We now restrict ourselves to $m \in \mathbf{N}$ and study these polynomials in more detail. The first few polynomials are given in the following table.

m	P_m^p	Q_m^p
0	1	1
1	$\frac{1}{p} - y$	$\frac{1}{2}(1 + \frac{1}{p} - y)$
2	$\frac{1}{2}\left[\left(y - \frac{1}{p}\right)^2 + \frac{1}{p}\right]$	$\frac{1}{3}\left[y + \frac{1}{2}(1 + \frac{1}{p} - y)(2 + \frac{1}{p} - y)\right]$

The following useful results, which hold for $m \geq 1$, are easily checked by induction. $B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$ is the beta function.

$$P_m^p(0) = \frac{\Gamma(m + \frac{1}{p})}{\Gamma(m+1)\Gamma(\frac{1}{p})} = \frac{1}{m\ B(m, \frac{1}{p})},$$
$$Q_m^p(0) = \frac{\Gamma(m + 1 + \frac{1}{p})}{\Gamma(m+2)\Gamma(\frac{1}{p})} = \frac{1}{(m+1)\ B(m+1, \frac{1}{p})},$$

$$P_m^p(y) = \frac{1}{m}\left[\frac{1}{p}\sum_{j=0}^{m-1} P_j^p(y) - y\ P_{m-1}^p(y)\right], \quad \text{and} \tag{36}$$

$$Q_m^p(y) = \frac{1}{m+1}\left[\frac{1}{p}\sum_{j=0}^{m-1} Q_j^p(y) - y\ Q_{m-1}^p(y) + 1\right]. \tag{37}$$

We now obtain two expressions for $\frac{d}{dx}V_m^p(x)$. First, observe that using (31) in (20) yields

$$\frac{d}{dx}V_m^p(x) = px^{p-1}\left(\left[P_m^p(x^p) - P_{m-1}^p(x^p)\right]V_0^p(x) + x\left[Q_{m-1}^p(x^p) - Q_{m-2}^p(x^p)\right]\right).$$

Differentiating (31) yields, after some simplifications,

$$\begin{aligned}\frac{d}{dx}V_m^p(x) = px^{p-1}\big[&(P_m^p)'(x^p)V_0^p(x) + P_m^p(x^p)\left[V_0^p(x) - V_{-1}^p(x)\right] \\ &+\ x^p(Q_{m-1}^p)'(x^p)V_{-1}^p(x) + \tfrac{1}{p}Q_{m-1}^p(x^p)V_{-1}^p(x)\big],\end{aligned}$$

where $(P_m^p)'(y)$ denotes $\frac{d}{dy}P_m^p(y)$. Equating these expressions yields

$$\begin{aligned}&-\left[(P_m^p)'(x^p) + P_{m-1}^p(x^p)\right]V_0^p(x) \\ &\quad = \left[x^p\left(Q_{m-1}^p\right)'(x^p) - P_m^p(x^p) + (\tfrac{1}{p} - x^p)Q_{m-1}^p(x^p) + x^pQ_{m-2}^p(x^p)\right]V_{-1}^p(x).\end{aligned} \tag{38}$$

This provides motivation for the following lemma.

LEMMA 13. *For $m \in \mathbf{N}$, $m \geq 1$,*

$$\frac{d}{dy}P_m^p(y) = -P_{m-1}^p(y) \quad \textit{and} \tag{39}$$

$$y\ \frac{d}{dy}Q_{m-1}^p(y) = P_m^p(y) - (m+1)Q_m^p(y) + mQ_{m-1}^p(y). \tag{40}$$

Proof. We first prove (39) by induction. It can be verified for $m = 1, 2$ using the table above. Then using (34), we find

$$\begin{aligned} & m\frac{d}{dy}P_m^p(y) \\ &= \left(m - 1 + \tfrac{1}{p} - y\right)\frac{d}{dy}P_{m-1}^p(y) - P_{m-1}^p(y) + y\frac{d}{dy}P_{m-2}^p(y) + P_{m-2}^p(y) \\ &= -P_{m-1}^p(y) - \left(m - 2 + \tfrac{1}{p} - y\right)P_{m-2}^p(y) - y\ P_{m-3}^p(y) \\ &= -mP_{m-1}^p(y). \end{aligned}$$

This implies that the coefficient of V_0^p in (38) is identically zero. Therefore the coefficient of V_{-1}^p must also be identically zero. Substituting $y = x^p$ gives

$$\begin{aligned} y\ \frac{d}{dy}Q_{m-1}^p(y) &= P_m^p(y) + \left(y - \frac{1}{p}\right)Q_{m-1}^p(y) - yQ_{m-2}^p(y) \\ &= P_m^p(y) - (m+1)Q_m^p(y) + mQ_{m-1}^p(y), \end{aligned}$$

where we used (35). □

Note that since the left-hand side of (40) is a polynomial of degree $m - 1$, this implies the coefficients of the y^m terms in P_m^p and $(m+1)Q_m^p(y)$ are identical. In fact, one can use (34) and (35) to see that the leading terms of P_m^p is $(-1)^m y^m/m!$ and that of Q_m^p is $(-1)^m y^m/(m+1)!$.

A set of polynomials $\{p_n(x)\}$belongs to the class known as Appell polynomials [4] if they satisfy $\frac{d}{dx}p_n(x) = p_{n-1}(x)$. Therefore, (39) implies that for each fixed p, the set $(-)^m P_m^p(y)$ forms a family of Appell polynomials.

One can use (39) in (34) to replace P_{m-1}^p and P_{m-2}^p by derivatives of P_m^p and obtain a second order differential equation satisfied by P_m^p. This allows us to obtain a relationship between the polynomials P_m^p and confluent hypergeometric functions, which we denote by ${}_1F_1(\alpha, \gamma, y)$.

THEOREM 14. *For $m \in \mathbf{N}$, $m \geq 1$, $P_m^p(y)$ satisfies the differential equation*

$$y\phi''(y) - \left(m - 1 + \frac{1}{p} - y\right)\phi'(y) - m\phi(y) = 0. \tag{41}$$

Standard techniques show that (41) has a polynomial solution of the form $\phi(y) = \sum_{k=0}^m b_k y^k$ with $b_k = -\frac{m+1-k}{k(m+\frac{1}{p}-k)}b_{k-1}$, $k \geq 1$, $b_0 \neq 0$ arbitrary, and a second solution of the form $\phi(y) = \sum_{k=0}^\infty c_k y^{k+m+1/p}$ with $c_k = -\frac{k-1+\frac{1}{p}}{k(m+\frac{1}{p}+k)}c_{k-1}$, $k \geq 1$, $c_0 \neq 0$ arbitrary. Since $P_m^p(0) = \frac{1}{m\ B(m,\frac{1}{p})}$, we conclude that $b_k = \frac{(-1)^k}{k!\ (m-k)B(m-k,\frac{1}{p})}$ and

$$P_m^p(y) = \sum_{k=0}^m \frac{(-1)^k\Gamma(m + \frac{1}{p} - k)}{\Gamma(k+1)\Gamma(m+1-k)\Gamma(\frac{1}{p})}\ y^k.$$

The restriction that $1/p$ be noninteger in the next result is neither serious nor unexpected in view of Proposition 3.

COROLLARY 15. *Let $p \neq \frac{1}{n}$ for $n \in \mathbf{N}$. Then*

$$P_m^p(y) = \frac{1}{m\ B(m, \frac{1}{p})}\ e^{-y}\ {}_1F_1\left(1 - \tfrac{1}{p}, 1 - \tfrac{1}{p} - m, y\right). \tag{42}$$

Proof. Write $\phi(y) = e^{-y}\widehat{\phi}(y)$. Then it follows from (41) that $\widehat{\phi}$ satisfies

$$y\widehat{\phi}''(y) - \left(m - 1 + \frac{1}{p} + y\right)\widehat{\phi}'(y) - \left(1 - \frac{1}{p}\right)\widehat{\phi}(y) = 0, \tag{43}$$

which has the form of the differential equation satisfied by the confluent hypergeometric function. Comparing the behavior of $P_m^p(y)$ near $y = 0$ with that of the well-known solutions to (43) suffices to complete the proof. ∎

It is well known [3] that for real α and γ, ${}_1F_1(\alpha, \gamma, y)$ has at most finitely many zeros on the real line. Hence, the same holds for P_m^p. In fact, we can show that P_m^p has no zeros when m is even and exactly one when m is odd.

To show this, it is convenient to introduce the new variable $z = \frac{1}{p} - y$ and write $P_m^p(y) = \tilde{P}_m^p(\frac{1}{p} - y)$. The first few of these polynomials are $\tilde{P}_0^p(z) = 1$, $\tilde{P}_1^p(z) = z$, and $\tilde{P}_2^p(z) = \frac{1}{2}(z^2 + \frac{1}{p})$.

LEMMA 16. *For $m \in \mathbf{N}$, $m \geq 2$, the polynomials $\tilde{P}_m^p(z)$ satisfy*

(i) $\tilde{P}_m^p(z) = \frac{1}{m}[(m - 1 + z)\tilde{P}_{m-1}^p(z) + (\frac{1}{p} - z)\tilde{P}_{m-2}^p(z)]$,

(ii) $\tilde{P}_m^p(z) = \frac{1}{m}[\frac{1}{p}\sum_{j=0}^{m-2}\tilde{P}_j^p(z) + z\tilde{P}_{m-1}^p(z)]$, *and*

(iii) $\frac{d}{dz}\tilde{P}_m^p(z) = \tilde{P}_{m-1}^p(z)$.

Proof. The proof follows immediately from substitution in (34), (36), and (39).

COROLLARY 17. *For $m \in \mathbf{N}$, $m \geq 0$, all coefficients in the polynomials $\tilde{P}_m^p(z)$ are positive.*

Proof. This follows immediately from the explicit expressions above for $\tilde{P}_m^p$ when $m = 0, 1$ and part (ii) of Lemma 16.

PROPOSITION 18. *Let $m \in \mathbf{N}$, $m \geq 1$.*

(i) *If m is even, $\tilde{P}_m^p(z) \geq 0$.*

(ii) *If m is odd, $\tilde{P}_m^p(z)$ has exactly one root z_m.*

Moreover, the roots form a strictly decreasing sequence with $-m + 1 \leq z_m \leq 0$.

Proof. First note that Corollary 17 implies that $\tilde{P}_m^p(z) \geq 0$ for all $z \geq 0$ and $\lim_{z\to\infty}\tilde{P}_m^p(z) = \infty$. Now we claim that

$$\lim_{z\to-\infty}\tilde{P}_m^p(z) = \begin{cases} -\infty & \text{if } m \text{ is odd,} \\ \infty & \text{if } m \text{ is even.} \end{cases}$$

This can easily be verified by induction using part (i) of Lemma 16 above.

For m odd, $\tilde{P}_m^p(z)$ has at least one root z_m. We now prove by induction that z_m is the only root if m is odd and that $\tilde{P}_m^p(z) \geq 0$ for all z if m is even. The induction hypothesis is easily seen to hold for $m = 0, 1$. Suppose it is true up to $m - 1$ and consider $\tilde{P}_m^p$, with m odd. Then, by Lemma 16(iii), $\frac{d}{dz}\tilde{P}_m^p(z) = \tilde{P}_{m-1}^p(z)$. Since $m - 1$ is even, by the induction hypothesis $\tilde{P}_{m-1}^p(z) \geq 0$. Thus $\tilde{P}_m^p(z)$ is increasing for all $z \in \mathbf{R}$, which implies that for m odd $\tilde{P}_m^p(z)$ has only one root. Since $\tilde{P}_m^p(z) > 0$ when $z > 0$, that one root must satisfy $z_m \leq 0$.

If m is even, then $\frac{d}{dz}\tilde{P}_m^p(z) = \tilde{P}_{m-1}^p(z)$ which, by the induction hypothesis, has exactly one root $z_{m-1} \leq 0$. Therefore $\tilde{P}_m^p(z)$ has a local extremum at z_{m-1}. Since $\frac{d^2}{(dz)^2}\tilde{P}_m^p(z) = \tilde{P}_{m-2}^p(z)$ and $m - 2$ is even, $\tilde{P}_{m-2}^p(z) \geq 0$ by the induction hypothesis and z_{m-1} is a local minimum for $\tilde{P}_m^p(z)$. By Lemma 16(i) we have

$$m\tilde{P}_m^p(z_{m-1}) = \left(\frac{1}{p} - z_{m-1}\right)\tilde{P}_{m-2}^p(z_{m-1})$$

since $\tilde{P}^p_{m-1}(z_{m-1}) = 0$. But by the induction hypothesis $z_{m-1} \le 0$ and $\tilde{P}^p_{m-2}(z_{m-1}) > 0$, so that $\tilde{P}^p_m(z_{m-1}) > 0$ as required.

It remains to be shown that the roots are decreasing and bounded below by $-(m-1)$. Both can be easily checked for $m = 1, 3$ and then proved by induction using Lemma 16(i). We now let m be odd. Since $\tilde{P}^p_m(z)$ is increasing, to show that $z_m > -m+1$, it suffices to show that $\tilde{P}^p_m(-m+1) < 0$. For $z = -m+1$ the recursion relation reduces to

$$m\tilde{P}^p_m(-m+1) = \left(\tfrac{1}{p} + m - 1\right) \tilde{P}^p_{m-2}(-m+1),$$

which is negative by the induction assumption hypothesis that $z_{m-2} \ge -m+3$. To show that $z_m < z_{m-2}$ it suffices to show that $\tilde{P}^p_m(z_{m-2}) > 0$. But

$$m\tilde{P}^p_m(z_{m-2}) = (m - 1 + z_{m-2})P^p_{m-1}(z_{m-2}) \ge 0$$

since $P^p_{m-2}(z_{m-2}) = 0$, $z_{m-2} > -m+3 > -m+1$, and $P^p_{m-1}(z)$ is positive. □

We now restate the results above in terms of the behavior of the original polynomials $P^p_m(y)$.

COROLLARY 19. *Let $m \in \mathbf{N}$, $m \ge 1$. Then*

(i) *if m is even, then $P^p_m(y) \ge 0$ for all $y \in \mathbf{R}$;*
(ii) *if m is odd, then $P^p_m(y)$ has exactly one root $y_m \ge 0$;*
(iii) *for all $m \ge 1$,*

$$\lim_{y\to-\infty} P^p_m(y) = \infty,$$

$$\lim_{y\to\infty} P^p_m(y) = \begin{cases} -\infty & \text{if } m \text{ is odd,} \\ \infty & \text{if } m \text{ is even.} \end{cases}$$

Although we were able to obtain an explicit expression for the polynomials $P^p_m(x)$ relating them to confluent hypergeometric functions and analyze their behavior in some detail, we do not have much information about $Q^p_m(x)$. This is, at least in part, because (40) mixes $Q^p_m(x)$ and $P^p_m(x)$ and does not lead directly to a differential equation for $Q^p_m(x)$. It would be interesting to know more about the polynomials $Q^p_m(x)$.

4. Inequalities and convexity.

4.1. Inequalities for $V_0(x)$. We first illustrate our strategy by proving a special class of inequalities for V_0. The convexity of $1/V_0(x)$ follows directly from the optimal upper bound in this class as given in Theorem 20 below. Although, as discussed at the end of this section, these inequalities generalize to V_m, the resulting upper bound is not sufficient to establish the convexity of $1/V_m$. For this we need a bound on the ratio $V_m(x)/V_{m-1}(x)$. Nevertheless, these simple inequalities for V_0, which can also be interpreted as ratio bounds, are of some interest in their own right in a variety of applications. Because the geometric strategy is also used in our more complex proofs of ratio bounds, we think there is some merit in presenting it first here.

We now define

$$g_k(x) = \frac{k}{(k-1)x + \sqrt{x^2 + k}}. \tag{44}$$

THEOREM 20. *For $x \ge 0$*

$$g_\pi(x) \le V_0(x) < g_4(x) \tag{45}$$

and these inequalities are optimal for functions of the form (44) *with equality only at* $g_\pi(0) = V_0(0) = \sqrt{\pi}$.

Proof. It is easy to see that the family of functions $g_k(x)$ is increasing in k and that $0 < g_k(x) < 1/x$. In order to prove that the upper bound is optimal, we first observe that $g_k'(x) = -k[g_k(x)]^2[x + (k-1)\sqrt{x^2+k}\,]/\sqrt{x^2+k}$ and $xg_k(x) - 1 = -kg_k(x)/[x + \sqrt{x^2+k}\,]$. Then one can verify that

$$\begin{aligned} g_k'(x) &> 2[xg_k(x) - 1] \\ &\iff \frac{k}{\sqrt{x^2+k}}\,\frac{(k-1)\sqrt{x^2+k}+x}{(k-1)x+\sqrt{x^2+k}} < \frac{2k}{x+\sqrt{x^2+k}} \\ &\iff (k-2)x^2 + k(k-3) < (k-2)x\sqrt{x^2+k} \\ &\iff x^2(k-2)(k-4) + k(k-3)^2 < 0 \end{aligned} \tag{46}$$

when $k > 3$. We now restrict attention to $3 \le k \le 4$ and let $h_k(x) = g_k(x) - V_0(x)$. For $k = 4$, the expression (46) implies $g_4'(x) < 2[xg_k(x) - 1]$ so that $h_4'(x) < 2xh_4(x)$ for all $x \ge 0$; whereas for $k < 4$ this holds only for $x < a_k = \sqrt{\frac{k(k-3)^2}{(k-2)(4-k)}}$. Since both $V_0(x)$ and $g_k(x)$ are positive and bounded above by $1/x$, their difference also satisfies $|h_k(x)| < 1/x \to 0$.

For $k = 4$, if $h_4(x) \le 0$ for some $x > 0$, then $h_4'(x) < 2xh_4(x)$ is negative and thus h_4 is negative and strictly decreasing from a certain x on, which contradicts $\lim_{x\to\infty} h_4(x) = 0$. Thus $h_4(x) > 0$ so that $g_4(x) > V_0(x)$ for all x. Now suppose that for some $k < 4$, g_k is an upper bound, i.e., $h_k(x) \ge 0$ for all $x \ge 0$. In particular, $h_k(x) \ge 0$ for all $x > a_k$. For $k < 4$, we find, however, that $h_k'(x) > 2xh_k(x)$ holds for $x > a_k$. Thus we get $h_k(x) \ge 0$ and strictly increasing for all $x > a_k$ which contradicts $\lim_{x\to\infty} h_k(x) = 0$. Thus the upper bound *cannot* hold when $x > a_k$ and $k < 4$. The lower bound also fails for $k > \pi$ since then $h_k(0) = g_k(0) - V_0(0) = \sqrt{k} - \sqrt{\pi} > 0$.

To establish the improved lower bound $g_\pi \le V_0(x)$, we note that the argument above implies that $h_k(x)$ is negative for $x > a_k$ and $3 < k \le \pi$. However, for $k < \pi$ we have $h_k(0) < 0$ so that $h_k(x)$ is also negative for very small x. If $h_k(x)$ is ever nonnegative, we can let b denote the first place $h_k(x)$ touches or crosses the x-axis, i.e., $h_k(b) = 0$ and $h_k(x) < 0$ for $x < b$. Then h_k must be increasing on some interval of the form (x_0, b). However, by the remarks above, $h_k(b) = 0$ implies $b \le a_k$ so that $h_k'(x) < 2xh_k(x) < 0$ on (x_0, b). Since this contradicts h_k increasing on (x_0, b), we must have $h_k(x) < 0$ for all $x \ge 0$ if $k < \pi$.

Thus we have proved the lower bound $g_k(x) < V_0(x)$ on $[0, \infty)$ for $k < \pi$. Since g_k is continuous and increasing in k, it follows that $g_\pi(x) \le V_0(x)$. To show that this inequality is strict except at $x = 0$, note that the right derivative of h_k at 0 satisfies $h_k'(0) = 2 - k$ so that $h_\pi'(0) < 0$ and $h_\pi(x)$ is negative at least on some small interval $(0, x_1)$. Then we can repeat the argument above to show that $h_\pi(x) < 0$ if $x > 0$. □

As discussed in [7, 18, 19] the upper bound implies the convexity of $1/V_0(x)$ on $(0, \infty)$; in fact, it is not hard to use the fact that (15) reduces to $\frac{d}{dx}V_0(x) = 2(xV_0 - 1)$ to see that the upper bound is equivalent to convexity. It was established independently by Wirth [19] and by Szarek and Werner [18]. (The latter actually proved slightly more by using (1) to define an asymmetric extension of $V_0(x)$ to negative x. They showed in [18] that this extension is convex for $x > -\frac{1}{\sqrt{2}}$.)

Both bounds in (45) are sharper than the inequalities of Komatsu [9, 15]. The weaker lower bound $g_3(x) < V_0(x)$ was used in [7] to show that the function

$[1/V_0(x) - x]^2 / V_0(x)$ is decreasing for $x \geq 0$. The lower bound $g_\pi(x) \leq V_0(x)$ was established earlier by Boyd [5] as the optimal bound in a different class of inequalities. There is an extensive literature (see, e.g., [15]) on bounds for $V_0(x)$; however, the class of inequalities obtained using functions of the form $g_k(x)$ does not seem to have been considered before so that the optimality of bounds of this type for $k = \pi$ and $k = 4$ seems new.

Mascioni [12] generalized the upper bound to $p \geq 2$ for which he showed

$$V_0^p(x) < \frac{4p}{3px^{p-1} + \sqrt{p^2x^{2p-2} + 8p(p-1)x^{p-2}}}$$

and also showed that this implies convexity of $1/V_0^p(x)$ for $p \geq 2$.

In view of property (a) of section 1.2, it would seem natural to try to generalize (20) using functions of the form

$$g_k^m(x) = \frac{k}{(k-1)x + \sqrt{x^2 + m + k}}. \tag{47}$$

Note that the functions g_k^m are increasing in k and that $\lim_{k\to\infty} g_k^m(x) = \frac{1}{x}$. Therefore property (a) implies that

$$g_1^m(x) \leq V_m(x) < \lim_{k\to\infty} g_k^m(x).$$

As g_k^m is continuous in k, there must exist i_m and j_m such that

$$g_{i_m}^m(x) \leq V_m(x) < g_{j_m}^m(x). \tag{48}$$

However, we have not obtained explicit expressions for i_m and j_m. One might expect that the optimal lower bound occurs when i_m is chosen to satisfy $g_{i_m}^m(0) = V_m(0)$. However, numerical evidence shows that this is false; in fact, this choice for i_m does not even yield an inequality.

4.2. Ratio bounds. One of our main goals is to show that the function $\frac{1}{V_m(x)}$ is convex for integer $m \geq 1$. The key to this is the realization that (45) can also be rewritten to give bounds on the ratio $V_0(x)/V_{-1}(x) = xV_0(x)$. We now let

$$G_k^m(y) = \frac{ky}{(k-1)y - m + \sqrt{(y+m)^2 + ky}} \tag{49}$$

and note that $xg_k(x) = G_k^0(x^2)$ so that (45) is equivalent to

$$G_\pi^0(x^2) \leq xV_0(x) = \frac{V_0(x)}{V_{-1}(x)} < G_4^0(x^2).$$

For integer $m > 0$, convexity of $\frac{1}{V_m(x)}$ can be shown to be equivalent to

$$R_m(x) \equiv \frac{V_m(x)}{V_{m-1}(x)} < G_4^m(x^2).$$

In addition to this upper bound, we can show the following theorem.

THEOREM 21. *Let $m \in \mathbf{N}$, $m \geq 0$. Then the inequalities*

$$G_8^{m-1}(x^2) < R_m(x) < G_4^m(x^2) \tag{50}$$

hold and are optimal in k for the class of functions of the form $G_k^m(x^2)$.

The upper bound is optimal in k for all m. The lower bound is optimal in the sense that 8 is the largest integer for which the lower bound in (50) holds for all m. However, as we discuss at the end of section 5.3, for fixed m one can find $k(m)$ such that $G^{m-1}_{k(m)}(x^2) < R_m(x)$ holds with $k(m) > 8$.

Since $G^m_k(y)$ is increasing in both m and k, its behavior at zero and infinity allows us to also draw some conclusions about the optimality in m of (50). $R_m(0) = 1 - \frac{1}{2m}$ and $G^m_k(0) = 1 - \frac{1}{1+2m}$ for all k. Therefore, $G^\nu_k(0) < R_m(0) < G^\mu_{k'}(0)$ implies $\nu \le m - \frac{1}{2}$ and $\mu \ge m - \frac{1}{2}$ for all k, k'. Thus, if we insist that m be integer, there is no choice of k which allows $m-1$ to be replaced by m in the lower bound when $m > 0$ or m by $m-1$ in the upper bound. This argument does not, however, rule out the possibility of bounds of the form $G^{m-1/2}_k(x^2) < R_m(x) < G^{m-1/2}_{k'}(x^2)$.

To examine the behavior at infinity, note that

$$G^m_k(y) = 1 - \frac{1}{2y} + \frac{4m+k+2}{8y^2} + O\left(\frac{1}{y^3}\right) \quad \text{and}$$
$$R_m(\sqrt{y}) = 1 - \frac{1}{2y} + \frac{4m+6}{8y^2} + O\left(\frac{1}{y^3}\right),$$

where the asymptotic expansion for R_m follows from Proposition 7. It then follows that $R_m(\sqrt{y}) < G^\mu_{k'}(y)$ implies $\mu > m+1-\frac{k}{4}$. Thus m is optimal for the upper bound if $k \le 4$ and any attempt to decrease m would require an increase in k. Furthermore, $\mu = m$ implies $k \ge 4$ so that the upper bound in (50) is optimal in k.

We postpone the proof of Theorem 21, which requires a lengthy computation even for the case $p = 2$, to the next section. Our proof uses induction on m. Therefore, we are able to establish (50) and the theorems in the next section only for m a positive integer. We believe that they are also true for noninteger m. However, a proof would require either a different method or independent verification of the upper bound for an initial range, such as $-1 < m < 0$.

The ratio $R_m(x)$ is of interest in its own right, and our results are sufficient to establish that it is increasing in x on $(0, \infty)$. This is proved in the next section after Theorem 23, which uses a similar argument.

THEOREM 22. *For $m \in \mathbf{N}$, the ratio $R_{m+1}(x) = \frac{V_{m+1}(x)}{V_m(x)}$ is increasing in x.*

4.3. Convexity of $1/V_m$. We now prove some important consequences of Theorem 21. The first is the following.

THEOREM 23. *For all $m \in \mathbf{N}$, the function $1/V_m(x)$ is convex on $[0, \infty)$.*

Proof. We need to show that

$$\left(\frac{1}{V_m(x)}\right)'' = \frac{2[V_m(x)']^2 - V_m(x)(V_m(x))''}{V_m(x)^3} > 0. \tag{51}$$

It follows from the differential equation (15) and the recursion relation (24) that

$$V_m(x)'' = 2\left[V_m(x)(1 + 2m + 2x^2) - 2V_{m-1}(x)(x^2 + m)\right]$$

so that (51) holds if and only if

$$[V_m(x)]^2(1 + 2m - 2x^2) + 2V_{m-1}(x)V_m(x)(3x^2 - m) - 4x^2[V_{m-1}(x)]^2 \le 0.$$

After division by $[V_{m-1}(x)]^2$ this can be rewritten as $P[R_m(x)] \le 0$, where

$$P(z) = z^2(1 + 2m - 2x^2) + 2z(3x^2 - m) - 4x^2.$$

Writing the roots of $P(z) = Az + 2Bz + C$ in the nonstandard form $\frac{-C}{B \pm \sqrt{B^2 - AC}}$, we find that $G_4^m(x^2)$ is either the smaller of two positive roots (when $x^2 > m + \frac{1}{2}$) or the only positive root (when $x^2 < m + \frac{1}{2}$). Since $P(0) < 0$ in both cases, we can conclude that

$$z < G_4^m(x^2) \quad \text{implies} \quad P(z) < 0.$$

Therefore, it follows from the upper bound in Theorem 21 that $P[R_m(x)] < 0$; hence (51) holds. □

Proof of Theorem 22. Using (15) one finds that

$$\frac{d}{dx} R_{m+1}(x) = 2x \left[\frac{R_{m+1}(x)}{R_m(x)} - 1 \right].$$

After rewriting this in terms of V_m and then using the recursion relation (24) with $p = 2$ to eliminate V_{m+1}, one finds that $R'_{m+1}(x) \geq 0$ if and only if

$$2(m+1)[R_m(x)]^2 - (2m + 1 - 2x^2)R_m(x) - 2x^2 \leq 0.$$

The polynomial $P(z) = 2(m+1)z^2 - (2m+1-2x^2)z - 2x^2$ has one positive and one negative root, and $R'_{m+1}(x) \geq 0$ if and only if $R_m(x)$ lies between these two roots. Since $1 \geq R_m(x) > 0$, it follows that $R_{m+1}(x)$ is increasing if and only if $R_m(x)$ is less than the larger root, i.e.,

$$R_m(x) \leq \frac{4x^2}{\sqrt{4(x^2+m)^2 + 1 + 4m + 12x^2} + 2x^2 - 2m - 1},$$

where we have again written the root in the nonstandard form $\frac{C}{-B + \sqrt{B^2 - AC}}$. Then using the upper bound of Theorem 21, we see that it suffices to show that

$$\begin{aligned} R_m(x) &\leq \frac{4x^2}{\sqrt{(x^2+m)^2 + 4x^2} + 3x^2 - m} \\ &\leq \frac{4x^2}{\sqrt{4(x^2+m)^2 + 1 + 4m + 12x^2} + 2x^2 - 2m - 1} \end{aligned}$$

or equivalently that

$$\begin{aligned} &\sqrt{(x^2+m)^2 + 4x^2} + 3x^2 - m \\ &\qquad \geq \sqrt{4(x^2+m)^2 + 1 + 4m + 12x^2} + 2x^2 - 2m - 1, \end{aligned}$$

which is easily checked.

5. Proof of ratio bounds. The proofs in this section, although elementary, are quite long and tedious. The details were checked using Mathematica.

5.1. Differential inequality. In order to prove Theorem 21, it suffices to establish the following.

LEMMA 24. *Let* G_k^m *be given by* (49). *Then*

(i) *for* $m \geq 1$, $\frac{d}{dx} G_4^m(x^2) \leq 2x\left(\frac{G_4^m(x^2)}{G_4^{m-1}(x^2)} - 1\right)$;

(ii) *for* $m \geq 4$, $\frac{d}{dx} G_8^m(x^2) \geq 2x\left(\frac{G_8^m(x^2)}{G_8^{m-1}(x^2)} - 1\right)$,

but the inequality (ii) *does not hold for* $m < 4$.

Proof. The proof is based on the elementary principle that if a function on the half-line is zero at the origin and increasing, then it is nonnegative. Unfortunately, the actual verification is rather tedious and requires the repeated use of this principle. For simplicity, we put $x^2 = y$ and assume $y \geq 0$. Then (i) is equivalent to

$$E_m(y) \equiv \left(\frac{G_4^m(y)}{G_4^{m-1}(y)} - 1 \right) - \frac{d}{dy} G_4^m(y) \geq 0. \tag{52}$$

Let $B_m = (m^2 + y^2 + 4y + 2my)^{\frac{1}{2}}$. Then

$$G_4^m(y) = \frac{4y}{B_m + 3y - m}$$

and

$$E_m(y) = \frac{B_m \left[4m + (B_m + 3y - m)(B_{m-1} - B_m + 1)\right] - (4m^2 + 8y + 4my)}{B_m \ (B_m + 3y - m)^2}.$$

Thus $E_m(y) \geq 0$ if and only if

$$\begin{aligned} &B_m(3m + B_{m-1}[B_m + 3y - m]) + m^3 + 2my \\ &\geq 3m^2 + 4y + 11y^2 + m^2 y + 5my^2 + 3y^3 + B_m((y+m)^2 + y). \end{aligned}$$

We put

$$\begin{aligned} s &= s(y,m) = 3m^2 + 4y + 11y^2 + m^2 y + 5my^2 + 3y^3, \\ t &= t(y,m) = 2my + m^3, \quad \text{and} \\ h &= h(y,m) = (y+m)^2 + y - 3m. \end{aligned}$$

Then $E_m(y) \geq 0$ if and only if

$$B_m B_{m-1}(B_m + 3y - m) \geq B_m h + s - t. \tag{53}$$

Notice that both sides of (53) are positive. For the left-hand side this follows immediately from $B_m > m$. For the right-hand side, note that $B_m h(0) + s(0) - t(0) = 0$ and

$$\begin{aligned} &\frac{d}{dy}\left[B_m h(y) + s(y)t(y)\right] \\ &= \frac{1}{B_m}\left[6y + 12my + 9m^2 y + 12y^2 + 9my^2 + 3y^3 \right. \\ &\qquad \left. + \ B_m(22y + 10my + 9y^2) + 4B_m - 2mB_m + 3m^3 - 6m + m^2 B_m\right]. \end{aligned}$$

Now observe that

$$\begin{aligned} &4B_m - 2mB_m + 3m^3 - 6m + m^2 B_m \\ &\qquad = 3m(m^2 - 2) + B_m(m^2 - 2m + 4) \geq 3m(m^2 - 1) \geq 0 \end{aligned}$$

since $m \geq 1$ and $B_m \geq m$. Hence $B_m h + s - t$ is increasing in y and the right-hand side of (53) is also positive. Therefore we can square both sides of (53) to conclude that it is equivalent to

$$F(y) = B_m f_1(y) - f_2(y) \geq 0, \tag{54}$$

where

$$f_1(y) = (m+y)(y^3 + my^2 + 3y^2 - m^2y - 3my + 2y - m^3 + 2m^2)$$

and

$$\begin{aligned} f_2(y) = {} & y^5 + 3my^4 + 5y^4 + 2m^2y^3 + 5my^3 + 2y^3 \\ & - 2m^3y^2 - 3m^2y^2 - 3m^4y - m^3y + 6m^2y - m^5 + 2m^4. \end{aligned}$$

Note that $F(0) = 0$. Therefore, to prove (54) it is enough to show that $\frac{d}{dy}F(y) = B_m f_1'(y) + \frac{f_1(y)(2+y+m)}{B_m} - f_2'(y) \geq 0$, or equivalently

$$D(y) \equiv d_1(y) - B_m d_2(y) \geq 0, \tag{55}$$

where

$$\begin{aligned} d_1(y) &= B_m^2 f_1'(y) + f_1(y)(2+y+m) \\ &= 6m^3 - m^4 - 3m^5 + 12my + 4m^2y - 13m^3y - 7m^4y + 20y^2 \\ &\quad + 14my^2 + 7m^2y^2 + 2m^3y^2 + 48y^3 + 49my^3 + 18m^2y^3 + 30y^4 \\ &\quad + 17my^4 + 5y^5 \quad \text{and} \\ d_2(y) &= f_2'(y) = 6m^2 - m^3 - 3m^4 - 6m^2y - 4m^3y \\ &\qquad + 6y^2 + 15my^2 + 6m^2y^2 + 20y^3 + 12my^3 + 5y^4. \end{aligned}$$

Note that $D(0) = 0$. Therefore, to prove (55) it is enough to show that $\frac{d}{dy}D(y) = d_1'(y) - B_m d_2'(y) - \frac{d_2(y)(2+y+m)}{B_m} \geq 0$, or equivalently

$$G(y) \equiv B_m g_1(y) - g_2(y) \geq 0, \tag{56}$$

where

$$\begin{aligned} g_1(y) &= d_1'(y) \\ &= 12m + 4m^2 - 13m^3 - 7m^4 + 40y + 28my + 14m^2y + 4m^3y \\ &\qquad + 144y^2 + 147my^2 + 54m^2y^2 + 120y^3 + 68my^3 + 25y^4 \quad \text{and} \\ g_2(y) &= B_m^2 d_2'(y) + d_2(y)(2+y+m) \\ &= 12m^2 + 4m^3 - 13m^4 - 7m^5 - 18m^2y - 13m^3y - 3m^4y \\ &\quad + 60y^2 + 180my^2 + 183m^2y^2 + 58m^3y^2 + 298y^3 \\ &\quad + 353my^3 + 122m^2y^3 + 170y^4 + 93my^4 + 25y^5. \end{aligned}$$

Note that $G(0) = 0$. Therefore, to prove (56) it is enough to show that $\frac{d}{dy}G(y) = B_m g_1'(y) + \frac{g_1(y)(2+y+m)}{B_m} - g_2'(y) \geq 0$, or equivalently

$$H(y) = h_1(y) - B_m h_2(y) \geq 0, \tag{57}$$

where

$$\begin{aligned} h_1(y) &= B_m^2 g_1'(y) + g_1(y)(2+y+m) \\ &= 24m + 60m^2 + 6m^3 - 13m^4 - 3m^5 + 240y + 300my + 460m^2y \\ &\qquad + 347m^3y + 113m^4y + 1520y^2 + 2246my^2 + 1663m^2y^2 + 482m^3y^2 \\ &\qquad + 2112y^3 + 2233my^3 + 738m^2y^3 + 930y^4 + 497my^4 + 125y^5 \quad \text{and} \\ h_2(y) &= g_2'(y) \\ &= -18m^2 - 13m^3 - 3m^4 + 120y + 360my + 366m^2y + 116m^3y \\ &\qquad + 894y^2 + 1059my^2 + 366m^2y^2 + 680y^3 + 372my^3 + 125y^4. \end{aligned}$$

Note that $H(0) = 12m(2 + 5m + 2m^2) > 0$. Therefore, to prove (57) it is enough to show that $\frac{d}{dy}H(y) = h_1'(y) - B_m h_2'(y) - \frac{h_2(y)(2+y+m)}{B_m} \geq 0$, or equivalently

$$l_1(y)B_m - l_2(y) \geq 0, \tag{58}$$

where

$$\begin{aligned}
l_1(y) &= h_1'(y) \\
&= 240 + 300m + 460m^2 + 347m^3 + 113m^4 + 3040y + 4492my \\
&\quad + 3326m^2y + 964m^3y + 6336y^2 + 6699my^2 + 2214m^2y^2 \\
&\quad + 3720y^3 + 1988my^3 + 625y^4 \quad \text{and} \\
l_2(y) &= B_m^2 h_2'(y) + h_2(y)(2 + y + m) \\
&= 84m^2 + 316m^3 + 347m^4 + 113m^5 + 720y + 2520my + 5046m^2y \\
&\quad + 3899m^3y + 1077m^4y + 9180y^2 + 15780my^2 + 11727m^2y^2 \\
&\quad + 3178m^3y^2 + 12202y^3 + 13145my^3 + 4202m^2y^3 + 4970y^4 \\
&\quad + 2613my^4 + 625y^5.
\end{aligned}$$

Note that $l_1(y) \geq 0$ and $l_2(y) \geq 0$ for all $y \geq 0$. Therefore (58) holds, if and only if $L(y) = B_m^2(l_1(y))^2 - (l_2(y))^2 \geq 0$, which follows immediately from the fact that all the coefficients are positive in

$$\begin{aligned}
L(y) = 4(&14400m^2 + 36000m^3 + 75936m^4 + 97368m^5 + 78972m^6 + 37188m^7 \\
&+ 8136m^8 + 57600y + 172800my + 717360m^2y + 1373400m^3y \\
&+ 1732428m^4y + 1360314m^5y + 599454m^6y + 111444m^7y + 1344000y^2 \\
&+ 3838560my^2 + 8437260m^2y^2 + 11062920m^3y^2 + 8495031m^4y^2 \\
&+ 3499083m^5y^2 + 595986m^6y^2 + 9342880y^3 + 24217360my^3 \\
&+ 32546720m^2y^3 + 24561680m^3y^3 + 9950080m^4y^3 + 1694280m^5y^3 \\
&+ 17918380y^4 + 37038224my^4 + 34271234m^2y^4 + 15627870m^3y^4 \\
&+ 2862630m^4y^4 + 15343236y^5 + 23700982my^5 + 13930330m^2y^5 \\
&+ 2982516m^3y^5 + 6445963y^6 + 6618363my^6 + 1887294m^2y^6 \\
&+ 1302640y^7 + 667056my^7 + 101250y^8).
\end{aligned}$$

To prove (ii) we proceed similarly, but now let $B_m(y) = \sqrt{(y+m)^2 + 8y}$ and $E_m(y) = \frac{d}{dy}G_8^m(y) - (\frac{G_8^m(y)}{G_8^{m-1}(y)} - 1)$, noting that

$$G_8^m(y) = \frac{8y}{B_m + 7y - m}.$$

We now need to show that $E_m(y) \geq 0$ for all $m \geq 4$. As the argument is similar to that above, we omit the details except to indicate the steps leading to the condition $m \geq 4$. Observe that $E_m(y) \geq 0$ if and only if

$$\begin{aligned}
&B_m((y+m)^2 + y - 7m) \\
&+ 7m^2 - m^3 + 24y - 2my + 5m^2y + 55y^2 + 13my^2 + 7y^3 \\
&\qquad \geq B_m B_{m-1}(7y - m + B_m).
\end{aligned} \tag{59}$$

Again, both sides of the inequality are positive. Hence we can square both sides of the inequality and, as above, get that (59) is equivalent to

$$F(y) = f_1(y) - B_m f_2(y) \geq 0, \tag{60}$$

with the appropriate f_1 and f_2. Again $F(0) = 0$. Therefore, in order to prove (60), it is enough to show that $\frac{d}{dy}F(y) \geq 0$, or equivalently, after rewriting,

$$D(y) = B_m d_1(y) - d_2(y) \geq 0, \tag{61}$$

with the appropriate d_1 and d_2. And again, $D(0) = 0$. We repeat the procedure: to prove (61), it is enough to show that $\frac{d}{dy}D(y) \geq 0$, or equivalently

$$e_1(y) - B_m e_2(y) \geq 0,$$

with the appropriate e_1 and e_2. Both e_1 and e_2 turn out to be positive for $y \geq 0$. Therefore (61) holds if

$$\begin{aligned} &L(y) = (e_1(y))^2 - (B_m e_2(y))^2 \geq 0, \\ &L(0) = 0, \text{ and} \\ &L'(0) = 192m^2(m-4)(1+2m)(480+64m+90m^2+33m^3). \end{aligned}$$

Thus $L'(0) \geq 0$ if and only if $m \geq 4$. For all $m \geq 4$, $L''(y) \geq 0$ for all $y \geq 0$. This finishes (ii). □

5.2. Proof of Theorem 21. We will prove by induction that $R_m(x) < G_4^m(x^2)$ for $m = 0, 1, 2, 3, \ldots$. As observed earlier, this inequality holds for $m = 0$, since it is then equivalent to the upper bound in (45). Let

$$H_m(x) = G_4^m(x^2) - R_m(x).$$

Then the upper bound in Theorem 21 is equivalent to $H_m(x) \geq 0$. This can be verified using the strategy of section 4.1 if the following conditions hold:

(i) $H_m(0) > 0$.
(ii) $\lim_{x\to\infty} H_m(x) = 0$.
(iii) $H_m'(x) \leq F_{\text{pos}}(x)H_m(x)$ for some strictly positive function $F_{\text{pos}}(x) > 0$.

Conditions (i) and (ii) hold. Indeed,

$$H_m(0) = \frac{2m}{1+2m} - \frac{\Gamma(m)\Gamma(m+\frac{1}{2})}{\Gamma(m+1)\Gamma(m-\frac{1}{2})} = \frac{1}{2m(1+2m)}$$

and

$$\lim_{x\to\infty} H_m(x) = 0,$$

since $\lim_{x\to\infty} R_m(x) = 1$, and for all $k \geq 1$, $\lim_{x\to\infty} G_k^m(x) = 1$.

We now check condition (iii). It follows from Lemma 24(i) and (52) that

$$\begin{aligned} H_m'(x) &\leq 2x\left[\frac{G_4^m(x^2)}{G_4^{m-1}(x^2)} - \frac{R_m(x)}{R_{m-1}(x)}\right] \\ &= \frac{2x}{G_4^{m-1}(x^2)R_{m-1}(x)}\left[G_4^m(x^2)R_{m-1}(x) - G_4^{m-1}(x^2)R_m(x)\right] \\ &\leq \frac{2x}{R^{m-1}(x)}\left[G_4^m(x^2) - R_m(x)\right] \\ &= \frac{2x}{R^{m-1}(x)}H_m(x), \end{aligned}$$

where the inequality follows from the induction hypothesis $R_{m-1}(x) < G_4^{m-1}(x^2)$. Thus (iii) holds with $F_{\rm pos}(x) = 2x/R_{m-1}(x)$.

For $m \geq 4$, the lower bound is proved similarly. One considers $H_m(x) = R_m(x) - G_8^m(x^2)$ instead and uses Lemma 24(ii). The cases R_1, R_2, and R_3 have to be verified directly as Lemma 24(ii) covers only the cases R_m for $m \geq 4$.

Using (24), $R_1 \geq G_8^0$ is equivalent to showing that

$$\frac{x}{V_0(x)} \geq \frac{9x + 14x^3 + (2x^2-1)\sqrt{8+x^2}}{2(7x + \sqrt{8+x^2})}.$$

As V_0 is always positive, this inequality holds trivially for those x for which the right-hand side is negative or zero. Therefore we need only to prove the inequality on the interval $[x_0, \infty)$, $x_0 \simeq 0.2511$, where

$$9x + 14x^3 + (2x^2-1)\sqrt{8+x^2} \geq 0.$$

Hence we need to show that for all $x \in [x_0, \infty)$,

$$\begin{aligned} V_0(x) &\leq 2x \frac{7x + \sqrt{8+x^2}}{9x + 14x^3 + (2x^2-1)\sqrt{8+x^2}} \\ &= 2x \frac{6x^2 - 1}{1 + 6x^2 + 12x^4 - 2x\sqrt{8+x^2}}. \end{aligned} \tag{62}$$

Put $h_1(x) = 2x \frac{6x^2-1}{1+6x^2+12x^4-2x\sqrt{8+x^2}}$. By Theorem 20 of section 4.1, inequality (62) is true for all $x \in [x_0, \infty)$, for which

$$g_4(x) \leq h_1(x).$$

This last inequality holds only on an interval $[x_0, x_1]$, $x_1 \simeq 1.399$. For all $x \geq x_1$, we show that

$$h_1(x) < \frac{1}{x}$$

and

$$h_1' \leq 2(xh_1 - 1).$$

Then (62) follows as in section 4.1.

Next, we find that $R_2 \geq G_8^1$ is equivalent to $V_0(x) \geq h_2(x)$, where

$$h_2(x) = 2x \frac{3 + 9x^2 + 14x^4 + (2x^2-3)(8x^2 + (1+x^2)^2)^{\frac{1}{2}}}{-3 - 7x^2 + 32x^4 + 28x^6 + (3 - 4x^2 + 4x^4)(8x^2 + (1+x^2)^2)^{\frac{1}{2}}}$$

and $R_3 \geq G_8^2$ is equivalent to $V_0(x) \leq h_3(x)$, where

$$h_3(x) = \frac{2x}{N(x)} \left[-30 - 23x^2 + 32x^4 + 28x^6 + \sqrt{8x^2 + (2+x^2)^2}\ (15 - 8x^2 + 4x^4) \right],$$

with

$$\begin{aligned} N(x) &= 30 + 3x^2 - 42x^4 + 92x^6 + 56x^8 \\ &+ (8x^2 + (2+x^2)^2)^{\frac{1}{2}}(-15 + 18x^2 - 12x^4 + 8x^6). \end{aligned}$$

Again, we have to check these inequalities for V_0 only for those x for which the right-hand sides are positive. We then proceed as for R_1 and show that $g_\pi \geq h_2$ up to a certain x_2 and that $h_2 < \frac{1}{x}$, $h_2' \geq 2(xh_1 - 1)$ on $[x_2, \infty)$. Similarly, we show that $g_4 \leq h_3$ up to a certain x_3 and that $h_3 < \frac{1}{x}$, $h_3' \leq 2(xh_1 - 1)$ on $[x_3, \infty)$. □

Note that these arguments also show that on the interval $[x_1, \infty)$ the function h_1 is a better upper bound for V_0 than g_4; on $[x_2, \infty)$ the function h_2 is a better lower bound for V_0 than g_π; and on $[x_3, \infty)$ the function h_3 is a better upper bound for V_0 than g_4. In fact, $h_3 \leq h_1 \leq g_4$ for $x > x_3$.

5.3. Optimality of bounds. We still need to consider optimality of the lower bound in upper bound in (50) in the parameter k. We continue the strategy above using similar notation so that now $B_m = \sqrt{(y+m)^2 + ky}$ and $E_m(y) = [\frac{G_k^m(y)}{G_k^{m-1}(y)} - 1] - \frac{d}{dy} G_k^m(y)$ with

$$G_k^m(y) = \frac{ky}{B_m + (k-1)y - m}.$$

Then $E_m(y) \leq 0$ if and only if

$$2B_m B_{m-1}(B_m + (k-1)y - m) \leq 2B_m(y + (y+m)^2 - (k-1)m) + P, \tag{63}$$

where

$$\begin{aligned} P = {} & -2m^2 + 2km^2 - 2m^3 - 2ky + k^2y - 4my - 6m^2y + 2km^2y \\ & - 2y^2 - 2ky^2 + 2k^2y^2 - 6my^2 + 4kmy^2 - 2y^3 + 2ky^3. \end{aligned}$$

For $m \geq 1$ and $k \geq 2$ both sides of the inequality are positive. Therefore we can square both sides and get that $E_m(y) \leq 0$ if and only if

$$F(y) = f_1(y) - B_m f_2(y) \geq 0,$$

where

$$\begin{aligned} f_1(y) = {} & -16km^4 + 8k^2m^4 + 32m^5 - 16km^5 - 20k^2m^2y + 8k^3m^2y + 8km^3y \\ & - 4k^2m^3y + 160m^4y - 96km^4y + 8k^2m^4y - 4k^3y^2 + k^4y^2 \\ & + 120km^2y^2 - 84k^2m^2y^2 + 12k^3m^2y^2 + 320m^3y^2 - 224km^3y^2 \\ & + 32k^2m^3y^2 + 20k^2y^3 - 20k^3y^3 + 4k^4y^3 + 152kmy^3 - 124k^2my^3 \\ & + 24k^3my^3 + 320m^2y^3 - 256km^2y^3 + 48k^2m^2y^3 + 56ky^4 - 52k^2y^4 \\ & + 12k^3y^4 + 160my^4 - 144kmy^4 + 32k^2my^4 + 32y^5 - 32ky^5 + 8k^2y^5, \\ f_2(y) = {} & 4\big(-4km^3 + 2k^2m^3 + 8m^4 - 4km^4 - 3k^2my + k^3my + 2km^2y \\ & - k^2m^2y + 32m^3y - 20km^3y + 2k^2m^3y + k^2y^2 + 16kmy^2 - 12k^2my^2 \\ & + 2k^3my^2 + 48m^2y^2 - 36km^2y^2 + 6k^2m^2y^2 + 10ky^3 - 9k^2y^3 + 2k^3y^3 \\ & + 32my^3 - 28kmy^3 + 6k^2my^3 + 8y^4 - 8ky^4 + 2k^2y^4\big). \end{aligned}$$

Then $F(0) = 0$ and in order that $F \geq 0$, we must have $\frac{d}{dy}F(0) \geq 0$. Computing $\frac{d}{dy}F(y)$, we find that $\frac{d}{dy}F(0) = 0$. We apply the same procedure as in the proof of Lemma 24, compute the successive derivatives, and evaluate them at 0. Evaluating the derivative at 0, in the fourth step of the procedure gives the value

$$24k^3m(1+2m)(km - 6m - k).$$

Therefore, in order that (63) (which is the condition for the lower bound) holds for all $m \geq 2$, we have to have at least that $k \geq \lim_{m\to\infty} \frac{6m}{m-1} = 6$. Thus for $m = 2$, $k \geq 12$ will do, for $m = 3$, $k \geq 9$, for $m = 4$, $k \geq 8$, and so forth. Therefore, as G_k^m is increasing in k, it seems a natural choice to pick $k = 12$ or bigger for the lower bound. And indeed, one can check that G_k^m satisfies the lower bound condition of Lemma 24 for $k \geq 12$ and $m \geq 2$. However, it is not true that for $k > 8$, G_k^{m-1} is a lower bound for R_m for *all* $m \geq 1$. It is a lower bound for all $m \geq m(k)$, from a certain $m(k)$ on. Thus the induction in the proof of Theorem 21 cannot start at $m = 0$ or $m = 1$. For $m < m(k)$, there exists x_m such that $R_m - G_k^{m-1} \geq 0$ on $[0, x_m]$ and $R_m - G_k^{m-1} < 0$ on (x_m, ∞).

5.4. Extensions to general *p*. For $p = 1$, all the functions involved are identically equal to 1 and hence trivially convex. For large x, $1/V_m^p(x) \approx x^{p-1}$ and x^{p-1} is concave for $1 < p < 2$. Hence we cannot expect convexity of $1/V_m^p$ on $(0, \infty)$ for p in $(1, 2)$. It was shown in [12] that $\frac{1}{V_0^p}$ is not convex on $\mathbf{R}^+$ for $0 < p < 1$. Therefore, we study only generalizations to $p > 2$. Our method of proof yields verification of the convexity of $\frac{1}{V_m^p}$ for all $m \geq 1$ up to at least $p = 4$. However, this method breaks down for larger p.

We generalize our previous notation to $R_m^p(x) = \frac{V_m^p(x)}{V_{m-1}^p(x)}$ and observe that

$$\frac{d}{dx} R_m^p(x) = px^{p-1} \left[\frac{R_m^p(x)}{R_{m-1}^p(x)} - 1 \right]. \tag{64}$$

For $k \geq 1$, $m \geq 0$, $p \geq 1$ we generalize G_k^m to

$$G_k^{m,p}(x^p) = \frac{kpx^p}{p[(k-1)x^p - m] + \sqrt{p^2(x^p + m)^2 + 2kp(p-1)x^p}}.$$

The proofs of Theorems 21, 22, and 23 can be extended provided that the analogue of Lemma 24 holds. This is not the case for large p. However, although the generalization of the upper bound in Theorem 21 is a necessary and sufficient condition for convexity of $1/V_m^p(x)$, Lemma 24 is only a sufficient condition for Theorem 21. Indeed, we were able to establish the lower bound in Theorem 21 for $m = 1, 2, 3$ even though part (ii) of Lemma 24 does not hold for $m < 4$. Hence, the fact that Lemma 24 breaks down for large p does not preclude convexity of $1/V_m^p(x)$. On the contrary, numerical evidence suggests that $\frac{1}{V_m^p}$ is convex for all $p \geq 2$.

LEMMA 25. *For all* $4 \geq p \geq 2$, $m \geq 1$,

$$\frac{d}{dx} G_4^{m,p}(x^p) \leq px^{p-1} \left[\frac{G_4^{m,p}(x^p)}{G_4^{m-1,p}(x^p)} - 1 \right].$$

This is equivalent to

$$E_m^p = px^{p-1} \left[\frac{G_4^{m,p}(x^p)}{G_4^{m-1,p}(x^p)} - 1 \right] - \frac{d}{dx} G_4^{m,p}(x^p) \geq 0,$$

which allows us to make some remarks about the range of validity. Although Lemma 25 can probably be extended to some higher p, it does *not* hold for all p, m. On the contrary, for all $m \geq 1$ there exists $p(m)$ and an interval $(x_1^{p(m)}, x_2^{p(m)})$ such that $E_m^p < 0$ on that interval for all $p \geq p(m)$. For example, numerical results show that

for $m = 1$, an interval on which $E_m < 0$ exists when $p \geq 10$; for $m = 2$ when $p \geq 14$; and for $m = 3$ when $p \geq 18$.

For simplicity, we sketch only the proof of Lemma 25 and give the final expressions for $p = 3$. Similar expressions can be given for $p = 4$. We have checked the details using Mathematica but omit the long formulas. As the expressions involved are monotone in p, this suffices for the entire interval $2 \leq p \leq 4$.

Proof. Let $B_m = \sqrt{p^2(y+m)^2 + 8p(p-1)y}$. Then $E_m^p \geq 0$ is equivalent to

$$\begin{aligned} &B_m B_{m-1}(B_m + 3py - pm) \\ &\quad \geq pB_m \left[p(y+m)^2 + y(5p-8) - 3pm\right] + p^2 \left\{3m^2p - m^3p \right. \\ &\qquad + \left. y(-8 + 8m + 8p - 6mp + m^2p) + y^2(-24 + 23p + 5mp) + 3py^3\right\}. \end{aligned}$$

Following the procedure used to prove Lemma 24(i), we eventually find that it would suffice to show that

$$l_1(y)B_m - l_2(y) \geq 0, \tag{65}$$

where, for $p = 3$,

$$\begin{aligned} l_1(y) = 3(&5120 + 2880m + 13800m^2 + 11034m^3 + 3051m^4 + 169600y \\ &+ 199632my + 116820m^2y + 26028m^3y + 298296y^2 + 239706my^2 \\ &+ 59778m^2y^2 + 133920y^3 + 53676my^3 + 16875y^4) \qquad \text{and} \\ l_2(y) = 27(&-2048m + 192m^2 + 3448m^3 + 3678m^4 + 1017m^5 25600y \\ &+ 49920my + 76152m^2y + 45330m^3y + 9693m^4y + 200000y^2 \\ &+ 250152my^2 + 139266m^2y^2 + 28602m^3y^2 + 195880y^3 \\ &+ 157254my^3 + 37818m^2y^3 + 59640y^4 + 23517my^4 + 5625y^5). \end{aligned}$$

Both $l_1(y)$ and $l_2(y)$ in (65) are positive, and hence (65) is equivalent to

$$L = l_1(y)^2 B_m^2 - l_2(y)^2 \geq 0,$$

which can be verified for $m \geq 1$. □

Acknowledgments. It is a pleasure to thank Professor S. Kwapien for the argument leading to the upper bound in property (a), and Professors R. Askey and M. Ismail for helpful discussions about the properties of hypergeometric functions. We would also like to thank Professor V. Mascioni for providing a copy of [12] before publication, and Professor Fink for drawing our attention to reference [15].

REFERENCES

[1] G.E. Andrews, R. Askey, and R. Roy, *Special Functions*, Cambridge University Press, Cambridge, UK, 1999.

[2] Y. Avron, I. Herbst, and B. Simon, *Strongly bound states of hydrogen in intense magnetic field*, Phys. Rev. A (3), 20 (1979), pp. 2287–2296.

[3] H. Bateman, *Higher Transcendental Functions*, Vol. 1, McGraw-Hill, New York, 1953.

[4] R.B. Boas, Jr. and R.C. Buck, *Polynomial Expansions of Analytic Functions*, Academic Press, New York, 1964.

[5] A.V. Boyd, *Inequalities for the Mills' ratio*, Rep. Statist. Appl. Res. Un. Jap. Sci. Engrs., 6 (1959), pp. 44–46.

[6] A.V. Boyd, *Note on a paper by Uppuluri*, Pacific J. Math., 22 (1967), pp. 9–10.

[7] R. Brummelhuis and M.B. Ruskai, *A one-dimensional model for many-electron atoms in extremely strong magnetic fields: Maximum negative ionization*, J. Phys. A, 32 (1999), pp. 2567–2582.

[8] R. Brummelhuis, M.B. Ruskai, and E. Werner, *One dimensional regularizations of the Coulomb potential with applications to atoms in strong magnetic fields*, in Differential Equations and Mathematical Physics, G. Weinstein and Weikhard, K., eds., International Press, Cambridge, MA, 2000, pp. 43–51.

[9] K. Ito and H.P. McKean, *Diffusion Processes and Their Sample Paths*, Springer-Verlag, New York, 1965, p. 17.

[10] W. Gautschi, *Some elementary inequalities relating to the gamma and incomplete gamma function*, J. Math. Phys., 38 (1959), pp. 77–81.

[11] N.N. Lebedev, *Special Functions and Their Applications*, rev. English ed., Prentice-Hall, Englewood Cliffs, NJ, 1965.

[12] V. Mascioni, *A generalization of an inequality related to the error function*, Nieuw Arch. Wisk. (4), 17 (1999), pp. 373–378.

[13] E.H. Lieb, J.P. Solovej, and J. Yngvason, *Heavy atoms in the strong magnetic field of a neutron star*, Phys. Rev. Lett., 69 (1992), pp. 749–752.

[14] E.H. Lieb, J.P. Solovej, and J. Yngvason, *Asymptotics of heavy atoms in high magnetic fields* I: *Lowest Landau band regions*, Comm. Pure Appl. Math., 47 (1993), pp. 513–591.

[15] D.S. Mitrinović, *Analytic Inequalities*, Springer-Verlag, New York, 1970, section 2.26.

[16] M.B. Ruskai and E. Werner, *A Pair of Optimal Inequalities Related to the Error Function*, preprint 97-564 of the Texas Mathematical Physics Preprint Archive, http://www.ma.utexas.edu/mp_arc/.

[17] E. Stein, *Singular Integrals and Differentiability Properties of Functions*, Princeton University Press, Princeton, NJ, 1970.

[18] S.J. Szarek and E. Werner, *Confidence regions for means of multivariate normal distributions and a non-symmetric correlation inequality for Gaussian measure*, J. Multivariate Anal., 68 (1999), pp. 193–211.

[19] M. Wirth, *On considère la fonction de* $\mathbf{R}$ *dans* $\mathbf{R}$ *définе par* $f(x) = e^{-x^2/2}$*; démontrer que la fonction* g *de* $\mathbf{R}$ *dans* $\mathbf{R}$ *définе par* $g(x) = f(x)/\int_x^\infty f(t)dt$ *est convexe*, Revue de Mathématiques Spéciales, 104 (1993), pp. 187–188.

[20] L. Schiff and H. Snyder, *Theory of the quadratic Zeeman effect*, Phys. Rev., 55 (1939), pp. 59–63.

SIAM J. MATH. ANAL.
Vol. 32, No. 3, pp. 465–492

UNIQUENESS OF EQUILIBRIUM CONFIGURATIONS IN SOLID CRYSTALS*

WILFRID GANGBO† AND ROBERTO VAN DER PUTTEN‡

Abstract. In this article, under suitable assumptions, it is proved that $\inf_{\mathbf{u}\in\mathcal{U}_\Lambda} E[\mathbf{u}]$ is dual to $\sup_{(a,b)}\{\int_\Omega a(\mathbf{F}(\mathbf{x}))d\mathbf{x} + \int_\Lambda b(\mathbf{y})d\mathbf{y}\}$, where, $E[\mathbf{u}] := \int_\Omega (h(\det D\mathbf{u}) - \mathbf{F}\cdot\mathbf{u})d\mathbf{x}$. Here, the infimum is performed over $\mathcal{U}_\Lambda$, the set of all orientation-preserving deformations $\mathbf{u} \in C^1(\Omega)^d$ that are homeomorphisms from $\bar{\Omega}$ onto $\bar{\Lambda}$, and the supremum is performed over the set of all upper semicontinuous functions a, b such that $a(\mathbf{z}) + \alpha b(\mathbf{y}) \le h(\alpha) - \mathbf{y}\cdot\mathbf{z}$. This duality result turns out to be important in the study of existence and uniqueness of smooth minimizers of E. Note that $M \to h(\det M)$ is not coercive and thus direct methods of the calculus of variations don't apply here.

Key words. Monge–Kantorovich, rearrangement, duality, solid crystals

AMS subject classifications. 49J40, 28A50

PII. S0036141099356684

Introduction. The theory of duality, one of the main tools in the calculus of variations, is well developed within the context of convex variational problems of the form $\inf_{\mathcal{U}} \int_\Omega L(\mathbf{x}, \mathbf{u}(\mathbf{x}), D\mathbf{u}(\mathbf{x}))d\mathbf{x}$, where the real-valued function $M \to L(\mathbf{x}, \mathbf{u}, M)$ defined on the set $\mathbf{R}^{d\times d}$ of the $d \times d$ matrices is convex for each $\mathbf{x} \in \Omega$ and $\mathbf{u} \in \mathbf{R}^d$. We recall that in the particular case $L(\mathbf{x}, \mathbf{u}, M) = g(M) - \mathbf{F}(\mathbf{x})\cdot\mathbf{u}$, where g is convex and coercive, then the duality statement is as follows: the infimum

$$\inf\left\{\int_\Omega L(\mathbf{x}, \mathbf{u}(\mathbf{x}), D\mathbf{u}(\mathbf{x}))d\mathbf{x} : \mathbf{u} \in W_0^{1,p}(\Omega, \mathbf{R}^d)\right\}$$

and the supremum

$$\sup\left\{-\int_\Omega g^*(-\mathbf{p}(\mathbf{x}))d\mathbf{x} : \mathbf{p} \in L^q(\Omega, \mathbf{R}^{d\times d}),\ \operatorname{div}\mathbf{p} = \mathbf{F}\right\}$$

coincide, where g^* is the Legendre transform of g. Furthermore, the extremum is attained in both problems (see [10]). An important class of nonconvex functions that occur in nonlinear elasticity theory is the class of *polyconvex* functions. There is no available theory of duality for that class. Recall that a real-valued function W of $\mathbf{R}^{d\times d}$ into $\mathbf{R}\cup\{+\infty\}$ is said to be *polyconvex* if it can be written as a convex function of the minors of M (see [8]). In this paper we consider a special class of polyconvex functions of the form $L(\mathbf{x}, \mathbf{u}, M) := W(M) - \mathbf{F}(\mathbf{x})\cdot\mathbf{u}$ and introduce a maximization problem, dual to $\inf_{\mathcal{U}} \int_\Omega L(\mathbf{x}, \mathbf{u}, D\mathbf{u})d\mathbf{x}$. As an application we study stable configurations of solid crystals occupying a reference configuration Ω and subject to a body force $\mathbf{F}$. If the crystal undergoes a deformation represented by a map $\mathbf{u} : \Omega \to \mathbf{R}^d$, $d \ge 2$ (in general $d = 3$), then its total energy functional is

$$E[\mathbf{u}] := \int_\Omega (W(D\mathbf{u}) - \mathbf{F}\cdot\mathbf{u})d\mathbf{x},$$

*Received by the editors May 24, 1999; accepted for publication (in revised form) January 13, 2000; published electronically September 15, 2000.
http://www.siam.org/journals/sima/32-3/35668.html
†School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332 (gangbo@math.gatech.edu). This author was supported by NSF grants DMS-96-22734 and DMS-99-70520.
‡Dipartimento di Matematica, Universita di Genova, Genova 16146, Italy (vanderpu@dima.unige.it).

where W represents the *Helmholtz free energy density.* In the framework of the continuum theory proposed by Ericksen [11] and [12], which has stimulated a growing body of work (see [23], [22], [21], [20], [25], [29], [28], [27], [26]), W belongs to a class of energy density functions that are invariant under change of lattice basis and frame:

$$W(M) = W(QMH) \tag{1}$$

for all $M \in \mathbf{R}^{d\times d}$, all $Q \in \mathbf{R}^{d\times d}$ such that $Q^TQ = I$, $\det Q > 0$, and all $H \in \mathbf{Z}^{d\times d}$, $|\det H| = 1$. The class of the energy densities suggested by Ericksen contains those of the form

$$W(M) = h(\det M) \qquad (M \in \mathbf{R}^{d\times d}), \tag{2}$$

where h is a convex function. In fact, it was shown by Chipot and Kinderlehrer [7] and Fonseca [15] that if W is of the form (1), then its quasi-convex envelope QW is of the form (2). Let us point out that the class of functions in (1) does not fall in the updated class of energy density functions of solid crystals. However, for purely mathematical interest, in what follows we choose to study the case where W satisfies (1), $QW = W$, and we still interpret the functional E as a solid crystal energy functional.

Following previous works (see, for instance, [17]) we assume that

$$h \in C^2(0, +\infty) \text{ is strictly convex,} \tag{3}$$

$$h(t) \to +\infty \text{ as } t \to 0^+ \text{ and } h(t)/t \to +\infty \text{ as } t \to +\infty. \tag{4}$$

We extend h to $\mathbf{R}$ by setting

$$h(t) := +\infty \text{ if } t \le 0. \tag{5}$$

Requirements (4) and (5) are imposed to make it energically impossible to compress part of the body of the crystal to zero volume, to extend part of the body excessively, or to change orientation. A typical example of body force is the gravity $\mathbf{F} = -g\,\mathbf{e}_d$, which can be written as the L^1-limit of a sequence of diffeomorphisms. Here we have set $\mathbf{e}_d := (0, \dots, 0, 1)$.

If the crystal undergoes a deformation $\bar{\mathbf{u}}$ under the action of the body force $\mathbf{F}$, then

$$-\operatorname{div}(\sigma_{\bar{\mathbf{u}}}) = \mathbf{F} \quad \text{in} \quad \Omega, \tag{6}$$

where $\sigma_{\bar{\mathbf{u}}}$ is the stress tensor $\frac{\partial W}{\partial M}(D\bar{\mathbf{u}})$. Solutions of (6) could be interpreted as critical points of the functional E.

A problem of great interest in nonlinear elasticity is the so-called *pure displacement boundary value problem:* given a diffeomorphism $\mathbf{u}_o$ from $\bar{\Omega}$ onto $\bar{\Lambda}$, where $\Lambda \subset \mathbf{R}^d$ is an open, bounded set, find $\bar{\mathbf{u}}$ *stable solution* of (6) such that the restrictions of $\bar{\mathbf{u}}$ and $\mathbf{u}_o$ on $\partial\Omega$ coincide. Stability means that not only is $\bar{\mathbf{u}}$ a critical point of E, but $\bar{\mathbf{u}}$ minimizes E over $\mathcal{U}_o$, the set of all maps $\mathbf{u}$ from $\bar{\Omega}$ onto $\bar{\Lambda}$ that are in $C^1(\Omega)^d$, $\det D\mathbf{u} > 0$, and such that the restrictions of $\mathbf{u}$ and $\mathbf{u}_o$ on $\partial\Omega$ coincide. Since $M \to h(\det M)$ is not coercive, and $\mathcal{U}_o$ is not closed under the weak topology on L^p spaces, the problem of minimizing E over $\mathcal{U}_o$ escapes the classical methods of the calculus of variations, and there is currently a wide literature on the subject. When $\mathbf{u}_o$ is the identity map and $\mathbf{F} = -g\,\mathbf{e}_d$ is the gravity force, Fonseca and Tartar [17]

showed that E has infinitely many minimizers in the set of displacements that are in $W^{1,\infty}(\Omega)^d$. Also, Chipot and Kinderlehrer [7] proved for E existence of parametrized measure minimizers by enlarging the set $\mathcal{U}_o$ to a set of Radon measures. We show that if $\mathbf{F} \in C^1(\bar{\Omega})^d$ is a homeomorphism, such that $\det D\mathbf{F} \in C^1(\bar{\Omega})^d$, $\det D\mathbf{F} > 0$, if Λ and $\mathbf{F}(\Omega)$ are convex, then the infimum

$$\inf_{\mathcal{U}_o} E \tag{7}$$

coincides with the infimum

$$\inf_{\mathcal{U}_\Lambda} E \tag{8}$$

and (8) admits a unique minimizer. Here, $\mathcal{U}_\Lambda$ is the set of all orientation-preserving maps $\mathbf{u} \in C^1(\Omega)^d$ that are homeomorphisms from $\bar{\Omega}$ onto $\bar{\Lambda}$.

One can interpret (8) as finding $\bar{\mathbf{u}}$ stable solution of the equations

$$\begin{cases} -\operatorname{div}\left[\frac{\partial W}{\partial M}(D\bar{\mathbf{u}})\right] &= \mathbf{F} \quad \text{in} \quad \Omega, \\ \bar{\mathbf{u}}(\Omega) &= \Lambda. \end{cases} \tag{9}$$

Uniqueness of a minimizer in (8) clearly implies that, in general, (7) does not admit a minimizer. In fact, sharper conclusions hold for a relaxation of (8): we substitute $\mathcal{U}_\Lambda$ by a bigger set $\mathcal{U}'_\Lambda$ containing maps which may not be smooth. We define $\mathcal{U}'_\Lambda$ to be the set of all maps $\mathbf{u}$ from Ω onto Λ that are one-to-one almost everywhere and such that $|\det D\mathbf{u}| \neq 0$ almost everywhere in the weak sense. Since it is delicate to define determinants of maps $\mathbf{u} \in \mathcal{U}'_\Lambda$ we define absolute values of determinants of these maps in the weak sense (see Definition 1.3). We denote by I the extension of $-E$ to $\mathcal{U}'_\Lambda$. In this new setting, under the assumptions that Ω, Λ, are bounded sets and $\mathbf{F} \in L^1(\Omega)^d$ is one-to-one, $(d-1)$-nondegenerate (see Definition 1.2), we prove that the following problem admits a unique maximizer

$$\sup_{\mathcal{U}'_\Lambda} I[\mathbf{u}], \tag{10}$$

where

$$I[\mathbf{u}] := \int_\Omega (\mathbf{F} \cdot \mathbf{u} - h(|\det D\mathbf{u}|))d\mathbf{x}.$$

If $\bar{\mathbf{u}}$ is the unique maximizer in (10), even if we drop the assumption that $\mathbf{F}$ is $(d-1)$-nondegenerate, then there exists a convex function $\psi_o : \mathbf{R}^d \to \mathbf{R}$ such that $\mathbf{F} = D\psi_o^* \circ \bar{\mathbf{u}}$, and

$$H(|\det D\bar{\mathbf{u}}|) = \psi_o^* \circ \bar{\mathbf{u}}. \tag{11}$$

Here

$$H(t) = h(t) - th'(t) \qquad (t \in \mathbf{R}),$$

and ψ_o^* is for the Legendre transform of ψ_o. One can readily check that

$$H \text{ is decreasing and } H(0, +\infty) = \mathbf{R}, \tag{12}$$

and so, if H^{-1} is of class C^1, smoothness of $|\det D\bar{\mathbf{u}}|$ is a straightforward consequence of (11). To understand the relation $\mathbf{F} = D\psi_o^* \circ \bar{\mathbf{u}}$, one can divide the computation of

the supremum in (10) into two steps. First, for each function $\alpha > 0$, we maximize $\mathbf{u} \to \int_\Omega \mathbf{F} \cdot \mathbf{u} d\mathbf{x}$ over the set of all $\mathbf{u}$ such that $\mathbf{u}(\Omega) = \Lambda$ and $|\det D\mathbf{u}| = \alpha$. Note that this intermediary variational problem is a Monge problem (see [3] and [18] in the case where $\alpha \equiv \chi_\Omega d\mathbf{x}$), and so the supremum is obtained for a map $\mathbf{u}^\alpha$ of the form $D\psi^\alpha \circ \mathbf{F}$, where ψ^α is a convex function. A sufficient condition for ψ^α to be differentiable at $\mathbf{F}(\mathbf{x})$ and thus for $D\psi^\alpha \circ \mathbf{F}$ to be well defined at $\mathbf{x}$ is that $\mathbf{F}$ be $(d-1)$-nondegenerate. Formally, if α_∞ maximizes the functional $\alpha \to \int_\Omega (\mathbf{F} \cdot D\psi^\alpha \circ \mathbf{F} - h(\alpha)) d\mathbf{x}$ over the set of all $\alpha > 0$, then $\bar{\mathbf{u}} = D\psi^{\alpha_\infty} \circ \mathbf{F}$ is a maximizer in (10).

Uniqueness of minimizers of E over $\mathcal{U}_\Lambda$ and $\mathcal{U}_o$ may clearly fail if we don't assume that $\mathbf{F}$ is $(d-1)$-nondegenerate. For instance, let $\mathbf{u}_o$ be the identity map, $\mathbf{F} \equiv 0$, and $h(t) = t^2/2 + 1/(2t^2)$. Since h attains it minimum for $t = 1$, any map $\mathbf{u} \in \mathcal{U}_o$ such that $\det D\mathbf{u} = 1$ is a minimizer of E over $\mathcal{U}_o$ and $\mathcal{U}_\Lambda$ where $\Lambda = \Omega$. Hence, E admits infinitely many minimizers over both sets $\mathcal{U}_o$ and $\mathcal{U}_\Lambda$. As shown in [17] it is necessary to have that $\det D\mathbf{F}(\mathbf{x}) \geq 0$ for E to admit a minimizer over $\mathcal{U}_o$.

Our primary and new contribution is to show that (10) is dual to the minimization problem (13):

$$\inf_{\mathcal{A}} J[\psi, \phi], \tag{13}$$

where

$$J[\psi, \phi] := \int_\Omega \psi(\mathbf{F}(\mathbf{x})) d\mathbf{x} + \int_\Lambda \phi(\mathbf{y}) d\mathbf{y}, \tag{14}$$

and $\mathcal{A}$ is the set of all pairs (ψ, ϕ) such that $\psi : \mathbf{R}^d \to \mathbf{R} \cup \{+\infty\}$ and $\phi : conv(\Lambda) \to \mathbf{R} \cup \{+\infty\}$ are lower semicontinuous, not identically $+\infty$, and

$$\psi(\mathbf{z}) + \alpha\phi(\mathbf{y}) + h(\alpha) \geq \mathbf{y} \cdot \mathbf{z}$$

for all $\mathbf{y} \in conv(\Lambda)$, all $\mathbf{z} \in \mathbf{R}^d$, and all $\alpha > 0$. To obtain the above duality result we first show that if μ is a finite positive measure on $\mathbf{R}^d$ of finite moments $M_o(\mu)$ and $M_1(\mu)$ (see (20)), then

$$\sup_{\gamma \in \Gamma(\mu)} \bar{I}[\gamma] = \inf_{(\psi,\phi) \in \mathcal{A}} J_\mu[\psi, \phi], \tag{15}$$

where

$$J_\mu[\psi, \phi] := \int_{\mathbf{R}^d} \psi(\mathbf{z}) d\mu(\mathbf{z}) + \int_\Lambda \phi(\mathbf{y}) d\mathbf{y},$$

and

$$\bar{I}[\gamma] := \int_C (\mathbf{y} \cdot \mathbf{z} - h(\alpha)) d\gamma(\alpha, \mathbf{y}, \mathbf{z}).$$

Here, $\Gamma[\mu]$ is the set of all Borel measures on $C := (0, +\infty) \times \mathbf{R}^d \times \mathbf{R}^d$ such that

$$\int_C f(\mathbf{z}) d\gamma(\alpha, \mathbf{y}, \mathbf{z}) = \int_{\mathbf{R}^d} f(\mathbf{z}) d\mu(\mathbf{z})$$

and

$$\int_C \alpha f(\mathbf{y}) d\gamma(\alpha, \mathbf{y}, \mathbf{z}) = \int_\Lambda f(\mathbf{y}) d\mathbf{y}$$

for all $f \in C_o(\mathbf{R}^d)$.

In fact, one can view $\Gamma(\mu)$ as a set containing $\mathcal{W}$, the set that consists of all Borel maps $\mathbf{w} : \mathbf{R}^d \to \Lambda$ such that the push forward of μ by $\mathbf{w}$ is absolutely continuous with respect to Lebesgue measure, say, $\mathbf{w}\sharp\mu = d\mathbf{y}/\beta(\mathbf{y})$ for some Borel function $\beta : \Lambda \to (0, +\infty)$. The inclusion $\mathcal{W} \subset \Gamma(\mu)$ means that we identify $\mathbf{w} \in \mathcal{W}$ to $\gamma^{\mathbf{w}} \in \Gamma(\mu)$, defined by

$$\int_C f(\alpha, \mathbf{y}, \mathbf{z}) d\gamma^{\mathbf{w}}(\alpha, \mathbf{y}, \mathbf{z}) := \int_{\mathbf{R}^d} f(\beta(\mathbf{w}(\mathbf{z})), \mathbf{w}(\mathbf{z}), \mathbf{z}) d\mu(\mathbf{z}) \tag{16}$$

for all $f \in C_o(\mathbf{R} \times \mathbf{R}^d \times \mathbf{R}^d)$. This definition makes sense provided that $\mathbf{w}$ is defined almost everywhere with respect to μ. Observe that if $\mu = \mu_{\mathbf{F}}$ where $\mu_{\mathbf{F}}[A] := |\mathbf{F}^{-1}[A]|$ is the d-dimensional Lebesgue measure of $\mathbf{F}^{-1}[A]$, then

$$\bar{I}[\gamma^{\mathbf{w}}] = I[\mathbf{w} \circ F]. \tag{17}$$

The plan is to first establish (15) and prove that the variational problems involved admit extremums under the general assumptions that h satisfies (3), (4), and (5) and that μ is a finite positive measure on $\mathbf{R}^d$ whose moments of order one are finite. Next we show that $\bar{I}$ admits a unique maximizer γ_o over $\Gamma(\mu)$. That maximizer can be parametrized over Λ: there is a map $\mathbf{m} : \Lambda \to \mathbf{R}^d$ and a function $\beta : \Lambda \to \mathbf{R}$, defined $\chi_\Lambda d\mathbf{y}$-almost everywhere such that

$$\int_C f(\alpha, \mathbf{y}, \mathbf{z}) d\gamma_o(\alpha, \mathbf{y}, \mathbf{z}) := \int_\Lambda f(\beta(\mathbf{y}), \mathbf{y}, \mathbf{m}(\mathbf{y})) d\mathbf{y}$$

for all $f \in C_o(\mathbf{R} \times \mathbf{R}^d \times \mathbf{R}^d)$. Then, we show that if $\mu_{\mathbf{F}}[A] := |\mathbf{F}^{-1}[A]|$ where $\mathbf{F}$ is one-to-one and $(d-1)$-nondegenerate, then every γ_o maximizing $\bar{I}$ over $\Gamma(\mu)$ is of the form $\gamma^{\mathbf{w}}$ (see (16)). Roughly speaking, $\mu[\mathbf{R}^d \setminus \mathbf{m}(\Lambda)] = 0$, $\mathbf{m}$ has an inverse $\mathbf{w}$ defined μ-almost everywhere. We combine (15) and (17) to deduce that $\mathbf{w} \circ F$ maximizes I, and that (10) is dual to (13). Simple examples such as $\mathbf{F}(\mathbf{x}) \equiv \mathbf{c}$ and $h(t) = t^2 + 1/t^2$ show that uniqueness of maximizer of $\bar{I}$ over $\Gamma(\mu)$ does not imply uniqueness of maximizer of I over $\mathcal{U}'_\Lambda$ unless the body force $\mathbf{F}$ is one-to-one and $(d-1)$-nondegenerate.

The remainder of the paper is organized as follows. In section 2 we prove existence of a minimizer (ψ_o, ϕ_o) of J_μ over $\mathcal{A}$ under the assumptions that h satisfies (3), (4), and (5) and that μ is a finite positive measure on $\mathbf{R}^d$ of finite moments $M_o(\mu)$ and $M_1(\mu)$. We write the Euler–Lagrange equations corresponding to the variational problem $\inf_{\mathcal{A}} J_\mu$ and deduce that if in addition μ vanishes on $(d-1)$-rectifiable subsets of $\mathbf{R}^d$, then there exist a convex function ψ and a positive Borel function β such that $D\psi\sharp\mu = d\mathbf{y}/\beta(\mathbf{y})$ and $\gamma_o = \gamma^{D\psi}$ maximizes $\bar{I}$ over $\Gamma(\mu)$. It is well known that a convex function is differentiable everywhere except on a $(d-1)$-rectifiable set (see [1]), and so the assumption that μ vanishes on $(d-1)$-rectifiable subsets of $\mathbf{R}^d$ is necessary to guarantee that $D\psi$ exists almost everywhere with respect to μ, so that the measure $\gamma_o = \gamma^{D\psi}$ be well-defined. Here, the analytical arguments used to write the Euler–Lagrange equations corresponding to $\inf_{\mathcal{A}} J_\mu$ are similar to the one independently introduced by Caffarelli–Varadhan [5] and the first author [18]. Having γ_o of the form $\gamma^{D\psi}$ readily yields that the duality (15) holds. By an approximation argument we extend (15) to the case where μ fails to vanish on $(d-1)$-rectifiable subsets of $\mathbf{R}^d$ and still obtain that supports of every maximizers of $\bar{I}$ over $\Gamma(\mu)$ are contained in the graph of a map from Λ into $(0, +\infty) \times \mathbf{R}^d$. We also show that the maximizer γ_o of $\bar{I}$ over $\Gamma(\mu)$ is unique.

In section 3, we assume that the given body force $\mathbf{F}$ belongs to $L^1(\Omega)$ and apply results of section 2 with $\mu[A] := |\mathbf{F}^{-1}[A]|$ to obtain that (10) is dual to (13). If in addition $\mathbf{F}$ is $(d-1)$-nondegenerate and one-to-one, then I admits a unique maximizer $\bar{\mathbf{u}}$ over $\mathcal{U}'_\Lambda$. Furthermore, $\bar{\mathbf{u}}$ satisfies $D\psi_o^* \circ \bar{\mathbf{u}} = \mathbf{F}$ and satisfies the Hamilton–Jacobi equation $H(|\det D\bar{\mathbf{u}}|) = \psi_o^* \circ \bar{\mathbf{u}}$ for some lower semicontinuous, convex function $\psi_o : \mathbf{R}^d \to \mathbf{R}$. Note that if $D\psi_o$ is differentiable almost everywhere with respect to μ, then we can conclude that $\bar{\mathbf{u}} = D\psi_o \circ \mathbf{F}$. Conversely, we show that if $\bar{\mathbf{u}} \in \mathcal{U}'_\Lambda$, $\psi_o : \mathbf{R}^d \to \mathbf{R}$ is a lower semicontinuous, convex function such that $H(|\det D\bar{\mathbf{u}}|) = \psi_o^* \circ \bar{\mathbf{u}}$ and $\mathbf{F} = D\psi_o^* \circ \bar{\mathbf{u}}$, then $\bar{\mathbf{u}}$ is the unique maximizer of I over $\mathcal{U}'_\Lambda$.

In section 4, using Caffarelli's regularity results on smoothness of convex potentials [4], [5], [6], we prove that if $\mathbf{F}$ and $\det D\mathbf{F}$ are of class C^1, if Λ and $\mathbf{F}(\Omega)$ are convex sets, then $\bar{\mathbf{u}}$ is of class C^1 and is the unique minimizer of E over $\mathcal{U}_\Lambda$. A corollary of this result is that given a diffeomorphism $\mathbf{u}_o$ of $\bar{\Omega}$ onto $\bar{\Lambda}$, the infima $\inf_{\mathcal{U}_\Lambda} E$ and $\inf_{\mathcal{U}_o} E$ coincide.

Four appendices are also provided. In Appendix A, we review basic facts about convex functions and study needed properties of the transformations introduced in Definition 1.6, $\phi \to \phi^\sharp$, $\psi \to \psi_\sharp$ from the set of real-valued functions to the set of convex functions. In Appendix C, we state that every one-to-one map $\mathbf{u} \in \mathcal{U}_\Lambda$ of class $C^1(\Omega) \cap C(\bar{\Omega})$ such that $\det D\mathbf{u} + \frac{1}{\det D\mathbf{u}}$ is bounded is a pointwise limit of a sequence of one-to-one maps $(\mathbf{u}_n) \subset \mathcal{U}_o$ of class $C^1(\Omega) \cap C(\bar{\Omega})$ with $\det D\mathbf{u}_n = \det D\mathbf{u}$. This approximation result is used in section 4 to prove that the infima $\inf_{\mathcal{U}_\Lambda} E$ and $\inf_{\mathcal{U}_o} E$ coincide. In Appendix D we recall facts on existence and smoothness of optimal maps in the Monge problem.

We next summarize the main results of the paper.

THEOREM 0.1 (main results). *Suppose that Ω, $\Lambda \subset \mathbf{R}^d$ are bounded open sets, that* (3), (4), *and* (5) *hold, and that $\mathbf{F} \in L^1(\Omega)^d$ is a Borel map. Then we have the following.*

(i) *Duality. J admits a minimizer (ψ_o, ϕ_o) over $\mathcal{A}$ and we have that $\inf_{\mathcal{A}} J[\psi, \phi] = \sup_{\mathcal{U}'_\Lambda} I[\mathbf{u}]$.*

(ii) *Uniqueness of a minimizer. If in addition $\mathbf{F}$ is one-to-one almost everywhere with respect to the d-dimensional Lebesgue measure and $|\mathbf{F}^{-1}(N)| = 0$ whenever N is $(d-1)$-rectifiable, then I admits a unique maximizer $\bar{\mathbf{u}}$ over $\mathcal{U}'_\Lambda$; we also have that $\bar{\mathbf{u}} = D\psi_o \circ \mathbf{F}$, and $H(|\det D\bar{\mathbf{u}}|) = \psi_o^* \circ \bar{\mathbf{u}}$, where (ψ_o, ϕ_o) minimizes J over $\mathcal{A}$.*

(iii) *Smoothness of the minimizer. Assume in addition that Ω is connected, its boundary $\partial\Omega$ is Lipschitz, and $\Lambda, \mathbf{F}(\Omega)$ are convex. If $\mathbf{F}$ and $\det D\mathbf{F}$ belong to $C^1(\bar{\Omega})^d$ and $\det D\mathbf{F} > 0$ on $\bar{\Omega}$, then $\bar{\mathbf{u}} \in \mathcal{U}_\Lambda \cap C^{0,s}(\bar{\Omega})^d$, $0 < \det D\bar{\mathbf{u}} \in C^{0,s}(\bar{\Omega}) \cap C^1(\Omega)$ for all $0 < s < 1$, $\bar{\mathbf{u}}$ is the unique minimizer of E over $\mathcal{U}_\Lambda$. Furthermore, we have that $-\operatorname{div}[\frac{\partial W}{\partial M}(D\bar{\mathbf{u}})] = \mathbf{F}$ in Ω in the weak sense.*

Proof. Parts (i) and (ii) follow from Theorem 3.1, and (iii) is a consequence of Theorem 4.1. □

Simple calculations show that the duality result obtained in Theorem 0.1 is

$$\inf_{\mathcal{U}'_\Lambda} \left\{ \int_\Omega (h(\det D\mathbf{u}) - \mathbf{F} \cdot \mathbf{u}) d\mathbf{x} \right\} = \sup_b \left\{ \int_\Omega L_b(\mathbf{F}(\mathbf{x})) d\mathbf{x} + \int_\Lambda b(\mathbf{y}) d\mathbf{y} \right\}, \tag{18}$$

where the supremum is performed over the set of all upper semicontinuous functions $b : \mathbf{R}^d \to \mathbf{R}$ and

$$L_b(\mathbf{z}) := \inf_{\mathbf{y} \in conv(\Lambda)} \{-\mathbf{y} \cdot \mathbf{z} - h^*(b(\mathbf{y}))\}.$$

1. Notations and definitions. For the convenience of the reader we collect together some of the notation introduced throughout the text.

• If $\Omega \subset \mathbf{R}^d$, then $\overline{\Omega}$ denotes the closure of Ω.

• B_R is the closed ball of center 0 and radius $R > 0$.

• $|A|$ stands for the d-dimensional Lebesgue measure of the set $A \subset \mathbf{R}^d$, and $\int_{\mathbf{R}^d} G d\mathbf{x}$ is the Lebesgue integral of G.

• If μ is a Borel measure on $\mathbf{R}^d$, then we denote by $\operatorname{spt}\mu$ the support of μ, which refers to the smallest closed set K such that $\mu[\mathbf{R}^d \setminus K] = 0$. If μ is absolutely continuous with respect to the d-dimensional Lebesgue measure and $\mu[A] = \int_A f d\mathbf{x}$ for $A \subset \mathbf{R}^d$ Borel, then we write $\mu = f d\mathbf{x}$.

• If μ is a Borel measure on $\mathbf{R}^d$ and $\mathbf{v} : \mathbf{R}^d \to \mathbf{R}^m$ is a Borel map, then we define $\mathbf{v}_{\#}\mu$ to be the Borel measure on $\mathbf{R}^m$ given by $\mathbf{v}_{\#}\mu[B] := \mu[\mathbf{v}^{-1}(B)]$ for $B \subset \mathbf{R}^m$.

• The characteristic function of $A \subset \mathbf{R}^d$ is denoted by χ_A.

• If $\psi : \mathbf{R}^d \to \mathbf{R} \cup \{+\infty\}$ is not identically $+\infty$, then the Legendre–Fenchel transform of ψ is the convex, lower semicontinuous function $\psi^* : \mathbf{R}^d \to \mathbf{R} \cup \{+\infty\}$ defined by

$$\psi^*(\mathbf{y}) := \sup_{\mathbf{x} \in \mathbf{R}^d} \{\mathbf{x} \cdot \mathbf{y} - \psi(\mathbf{x})\}. \tag{19}$$

• The subdifferential of a convex function $\psi : \mathbf{R}^d \to \mathbf{R} \cup \{+\infty\}$ is the set $\partial\psi \subset \mathbf{R}^d \times \mathbf{R}^d$ consisting of all $(\mathbf{x}, \mathbf{y})$ satisfying

$$\psi(\mathbf{z}) - \psi(\mathbf{x}) \geq \mathbf{y} \cdot (\mathbf{z} - \mathbf{x}) \qquad \text{for all } \mathbf{z} \in \mathbf{R}^d.$$

If $(\mathbf{x}, \mathbf{y}) \in \partial\psi$, we may also write $\mathbf{y} \in \partial\psi(\mathbf{x})$. Recall $\mathbf{x} \in \partial\psi^*(\mathbf{y})$ whenever $\mathbf{y} \in \partial\psi(\mathbf{x})$, while the converse also holds true if ψ is convex lower semicontinuous. In that case $\partial\psi$ is a closed set. In general, the set $\partial\psi(\mathbf{x}) \subset \mathbf{R}^d$ is closed and convex.

• **id** stands for the identity map $\mathbf{id}(\mathbf{x}) = \mathbf{x}$.

• We denote the set of all $d \times d$ matrices whose entries are real numbers by $\mathbf{R}^{d\times d}$.

• We denote the set of all homeomorphism from $A \subset \mathbf{R}^d$ onto $B \subset \mathbf{R}^d$ by $Diff^0(A, B)$. If $k \geq 1$ is an integer, $\Omega, \Lambda \subset \mathbf{R}^d$ are open, then $Diff^k(\Omega, \Lambda)$ is the set of all maps $\mathbf{v} \in Diff^0(\Omega, \Lambda)$ such that $\mathbf{v} \in C^k(\Omega)^d$ and $\mathbf{v}^{-1} \in C^k(\Lambda)^d$. We denote the set of all maps $\mathbf{v} \in Diff^0(\bar{\Omega}, \bar{\Lambda})$ such that $\mathbf{v}$ is of class C^k in a neighborhood of $\bar{\Omega}$ and $\mathbf{v}^{-1}$ is of class C^k in a neighborhood of $\bar{\Lambda}$ by $Diff^k(\bar{\Omega}, \bar{\Lambda})$.

• We define $\mathcal{U}_o$ to be the set of all continuous maps $\mathbf{u}$ from $\bar{\Omega}$ onto $\bar{\Lambda}$ that are in $C^1(\Omega)^d$, such that $\det D\mathbf{u} > 0$, $\mathbf{u}$, and $\mathbf{u}_o$ coincide on $\partial\Omega$. $\mathcal{U}_\Lambda$ is the set of all orientation-preserving maps $\mathbf{u} \in C^1(\Omega)^d$ that are homeomorphisms from $\bar{\Omega}$ onto $\bar{\Lambda}$. $\mathcal{U}'_\Lambda$ is the set of all maps $\mathbf{u}$ from Ω onto Λ that are one-to-one almost everywhere and such that $|\det D\mathbf{u}| \neq 0$ almost everywhere in the weak sense.

• We define $\mathcal{A}$ to be the set of all pairs of functions (ψ, ϕ) such that $\psi : \mathbf{R}^d \to \mathbf{R} \cup \{+\infty\}$, $\phi : conv(\Lambda) \to \mathbf{R} \cup \{+\infty\}$ are lower semicontinuous, not identically $+\infty$, and $\psi(\mathbf{z}) + \alpha\phi(\mathbf{y}) + h(\alpha) \geq \mathbf{y} \cdot \mathbf{z}$ for all $\mathbf{y} \in conv(\Lambda)$, $\mathbf{z} \in \mathbf{R}^d$, and all $\alpha > 0$.

We recall definitions needed in that which follows.

DEFINITION 1.1. *Let $A, B \subset \mathbf{R}^d$. We say that $\mathbf{v} : A \to B$ is* one-to-one almost everywhere *from A onto B (with respect to the d-dimensional Lebesgue measure) if $|B \setminus \mathbf{v}(A)| = 0$, if there exists a set $N \subset A$ such that $|N| = 0$, and if the restriction of $\mathbf{v}$ to $A \setminus N$ is one-to-one. By abuse of language we omit the expression "with respect to the d-dimensional Lebesgue measure."*

DEFINITION 1.2. *Let $A, B \subset \mathbf{R}^d$. We say that a Borel map $\mathbf{v} : A \to B$ is* nondegenerate *if $|\mathbf{v}^{-1}(N)| = 0$ whenever $|N| = 0$. We say that $\mathbf{v}$ is $(d-1)$-*nondegenerate *if $|\mathbf{v}^{-1}(N)| = 0$ whenever N is $(d-1)$-rectifiable.*

Recall that $N \subset \mathbf{R}^d$ is $(d-1)$-rectifiable if N is a countable union of $(d-1)$-hypersurfaces of class C^1, union a set of zero $(d-1)$-dimensional Hausdorff measure.

DEFINITION 1.3. *Let $A, B \subset \mathbf{R}^d$, and let $\beta_o \in L^1(A)$, $\beta_1 \in L^1(B)$ be nonnegative functions. Let $\mathbf{v} : A \to B$ be a* one-to-one almost everywhere *Borel map from A onto B. We say that $\beta_1(\mathbf{v}(\mathbf{x}))|\det D\mathbf{v}(\mathbf{x})| = \beta_o(\mathbf{x})$ in A* in the weak sense *if*

$$\int_A \varphi(\mathbf{v}(\mathbf{x}))\beta_o(\mathbf{x})d\mathbf{x} = \int_B \varphi(\mathbf{y})\beta_1(\mathbf{y})d\mathbf{y}$$

for all $\varphi \in C_o(\mathbf{R}^d)$.

Remark 1.4. Note that if $\mathbf{v}$ is *one-to-one almost everywhere*, and if $|\mathbf{v}^{-1}[C]| = |C|$ for every Borel set C, then $|\det D\mathbf{v}| = 1$ *in the weak sense* although $D\mathbf{v}$ may not exist.

DEFINITION 1.5. *Let μ and ν be two Borel measures on $\mathbf{R}^d$. We say that the Borel map $\mathbf{v} : \mathbf{R}^d \to \mathbf{R}^d$ pushes μ forward to ν and we write $\mathbf{v}\sharp\mu = \nu$ if $\mu[\mathbf{v}^{-1}(B)] = \nu[B]$ for all Borel sets $B \subset \mathbf{R}^d$.*

DEFINITION 1.6. *If ϕ and ψ are two real valued functions of subsets of $\mathbf{R}^d$ into $\mathbf{R} \cup \{+\infty\}$, then we define $\phi^\sharp$ and $\psi_\sharp$ to be the following convex functions of $\mathbf{R}^d$ into $\mathbf{R} \cup \{+\infty\}$:*

$$\phi^\sharp(\mathbf{z}) := \sup_{\mathbf{y} \in conv(\Lambda)} \{\mathbf{y} \cdot \mathbf{z} + h^*(-\phi(\mathbf{y}))\} \quad \text{and} \quad \psi_\sharp(\mathbf{y}) := \sup_{\alpha > 0} \left\{ \frac{\psi^*(\mathbf{y}) - h(\alpha)}{\alpha} \right\}.$$

2. An auxiliary variational problem: Duality. Throughout this section we assume that $\Lambda \subset \mathbf{R}^d$ is an open bounded set whose closure is contained in the closed ball B_{R_o} of center 0 and radius R_o. We assume that h satisfies (3), (4), (5) and μ is a finite positive measure on $\mathbf{R}^d$ of finite moments $M_o(\mu)$ and $M_1(\mu)$, where

$$M_o(\mu) := \mu[\mathbf{R}^d] < +\infty, \quad M_1(\mu) := \int_{\mathbf{R}^d} |\mathbf{z}| d\mu(\mathbf{z}) < +\infty. \tag{20}$$

We define

$$J_\mu[\psi, \phi] := \int_{\mathbf{R}^d} \psi(\mathbf{z})d\mu(\mathbf{z}) + \int_\Lambda \phi(\mathbf{y})d\mathbf{y}$$

and

$$\bar{I}[\gamma] := \int_C (\mathbf{y} \cdot \mathbf{z} - h(\alpha))d\gamma(\alpha, \mathbf{y}, \mathbf{z}),$$

where C is the set $(0, \infty) \times \mathbf{R}^d \times \mathbf{R}^d$. Let $\Gamma[\mu]$ be the set of all Borel measures on C such that

$$\int_C f(\mathbf{z})d\gamma(\alpha, \mathbf{y}, \mathbf{z}) = \int_{\mathbf{R}^d} f(\mathbf{z})d\mu(\mathbf{z})$$

and

$$\int_C \alpha f(\mathbf{y})d\gamma(\alpha, \mathbf{y}, \mathbf{z}) = \int_\Lambda f(\mathbf{y})d\mathbf{y}$$

for all $f \in C_o(\mathbf{R}^d)$. Observe that for every $(\psi, \phi) \in \mathcal{A}$ and every $\gamma \in \Gamma(\mu)$ we have that

$$J_\mu[\psi, \phi] = \int_C (\psi(\mathbf{z}) + \alpha\phi(\mathbf{y}))d\gamma \geq \int_C (\mathbf{y} \cdot \mathbf{z} - h(\alpha))d\gamma = \bar{I}[\gamma],$$

and so

$$\sup_{\Gamma(\mu)} \bar{I}[\gamma] \leq \inf_{\mathcal{A}} J_\mu[\psi, \phi]. \tag{21}$$

We establish the reverse inequality in this section.

Remark 2.1. Note that if $(\psi, \phi) \in \mathcal{A}$, then we have that

$$\psi^-(\mathbf{z}) \leq R_o|\mathbf{z}| + |h(1)| + \inf_{conv(\Lambda)} \phi^+ \tag{22}$$

for all $\mathbf{z} \in \mathbf{R}^d$ and

$$\phi^-(\mathbf{y}) \leq |\mathbf{y}||\mathbf{z}| + |h(1)| + \psi^+(\mathbf{z}) \tag{23}$$

for all $\mathbf{y} \in \text{conv}\,(\Lambda)$, $\mathbf{z} \in \mathbf{R}^d$. Combining (20), (22), and (23) we deduce that both $\int_{\mathbf{R}^d} \psi(\mathbf{z}) d\mu(\mathbf{z})$, $\int_\Lambda \phi(\mathbf{y}) d\mathbf{y}$ exist although they may be $+\infty$ and $J_\mu[\psi, \phi]$ is well-defined.

LEMMA 2.2. *The set $\mathcal{A}$ contains at least an element (ψ, ϕ). Also, there exists a constant c_a depending only on h, Λ, $M_o[\mu]$ such that*

(i) $|\inf_{\mathcal{A}} J_\mu| \leq c_a(1 + M_1[\mu])$;

(ii) *if ψ, ϕ are convex and $|J_\mu[\psi, \phi] - \inf_{\mathcal{A}} J_\mu| \leq 1$, then*

$$\int_\Lambda |\phi(\mathbf{y})| d\mathbf{y} \text{ and } \int_{\mathbf{R}^d} |\psi(\mathbf{z})| d\mu(\mathbf{z}) \leq c_a(1 + M_1[\mu]);$$

(iii) *if in addition $Lip(\psi) \leq R_o$, then we have that*

$$\psi(\mathbf{z}) \leq R_o|\mathbf{z}| + RR_o + \frac{c_a}{\mu[B_R]}(1 + M_1[\mu]) \qquad (\mathbf{z} \in \mathbf{R}^d).$$

Proof. Step 1. The set $\mathcal{A}$ is nonempty since it contains (ψ_o, ϕ_o), where $\phi_o(\mathbf{y}) := 1$ on $\text{conv}\,(\Lambda)$, $\psi_o(\mathbf{z}) := R_o|\mathbf{z}| - c$ on $\mathbf{R}^d$, and $c := \inf_{\alpha>0}\{h(\alpha) + \alpha\}$. We deduce that

$$\inf_{\mathcal{A}} J_\mu \leq J_\mu[\psi_o, \phi_o] \leq |\Lambda| + R_o M_1[\mu] - cM_o[\mu]. \tag{24}$$

If $(\psi, \phi) \in \mathcal{A}$, then

$$J_\mu[\psi, \phi] \geq -(\alpha\phi(\mathbf{y}_o) + h(\alpha))M_o[\mu] - R_o M_1[\mu] + \int_\Lambda \phi(\mathbf{y}) d\mathbf{y} \tag{25}$$

for all $\alpha > 0$ and all $\mathbf{y}_o \in \Lambda$. Setting $\alpha := |\Lambda|/M_o[\mu]$ in (25) and using (24) we have that

$$|\inf_{\mathcal{A}} J_\mu| \leq c_1, \tag{26}$$

where $c_1 := |\Lambda| + R_o M_1[\mu] + h(|\Lambda|/M_o[\mu])M_o[\mu]$.

Step 2. Let $(\psi, \phi) \in \mathcal{A}$ be such that $|J_\mu[\psi, \phi] - \inf_{\mathcal{A}} J_\mu| \leq 1$. In light of (25) we have that

$$\int_\Lambda \phi(\mathbf{y}) d\mathbf{y} \leq 1 + \inf_{\mathcal{A}} J_\mu + (\alpha\phi(\mathbf{y}_o) + h(\alpha))M_o[\mu] + R_o M_1[\mu] \tag{27}$$

for all $\alpha > 0$ and all $\mathbf{y}_o \in \Lambda$. Choosing α and $\mathbf{y}_o$ appropriately in (27) we have that

$$\left| \int_\Lambda \phi(\mathbf{y}) d\mathbf{y} \right| \leq c_2(1 + M_1[\mu]), \tag{28}$$

where c_2 is a constant depending only on h, Λ, $M_o[\mu]$. Combining (26) and (28) we deduce that there exists a constant c_3 depending only on h, Λ, and $M_o[\mu]$ such that

$$\left|\int_\Lambda \phi(\mathbf{y})d\mathbf{y}\right|, \quad \left|\int_{\mathbf{R}^d} \psi(\mathbf{z})d\mu(\mathbf{z})\right| \le c_3(1+M_1[\mu]). \tag{29}$$

Step 3. Assume that $(\psi,\phi)\in\mathcal{A}$, ϕ is convex on $conv(\Lambda)$, ψ is convex on $\mathbf{R}^d$, and $|J_\mu[\psi,\phi]-\inf_{\mathcal{A}} J_\mu|\le 1$. In light of (29) there exists $\mathbf{z}_o\in\mathbf{R}^d$ such that

$$|\psi(\mathbf{z}_o)|\le c_3(1+M_1[\mu])/\mu[\mathbf{R}^d]. \tag{30}$$

Integrating (23) over $\mathbf{R}^d$ we have that

$$M_o[\mu]\phi^-(\mathbf{y})\le|\mathbf{y}|M_1[\mu]+|h(1)|M_o[\mu]+\int_{\mathbf{R}^d}\psi^+(\mathbf{z})d\mu(\mathbf{z}) \tag{31}$$

for all $\mathbf{y}\in conv(\Lambda)$. Either $\inf_{conv(\Lambda)}\phi^+>0$, in which case

$$\phi^-\equiv 0 \qquad \text{on } conv(\Lambda), \tag{32}$$

or $\inf_{conv(\Lambda)}\phi^+=0$, in which case (22) and (29) imply that there exists a constant c_4 depending only on h, Λ, and $M_o[\mu]$ such that

$$\int_{\mathbf{R}^d}|\psi(\mathbf{z})|d\mu(\mathbf{z})\le c_4(1+M_1[\mu]),$$

which, combined with (31), yields

$$M_o[\mu]\phi^-(\mathbf{y})\le|\mathbf{y}|M_1[\mu]+|h(1)|M_o[\mu]+c_4(1+M_1[\mu]) \tag{33}$$

for all $\mathbf{y}\in conv(\Lambda)$. Using (32) and (33) we deduce that in any case, there exists a constant c_5 depending only on h, Λ, and $M_o[\mu]$ such that

$$\phi^-(\mathbf{y})\le c_5(1+M_1[\mu]) \tag{34}$$

for all $\mathbf{y}\in conv(\Lambda)$. In light of (29) and (34) we have that there exists a constant c_6 depending only on h, Λ, and $M_o[\mu]$ such that

$$\int_\Lambda|\phi(\mathbf{y})|d\mathbf{y}\le c_6(1+M_1[\mu]). \tag{35}$$

Since ϕ is convex, (35) implies that for each $K\subset\Lambda$ compact set, there exists a constant c_K depending only on h, Λ, $M_o[\mu]$, and K such that (see [13, p. 236])

$$|\phi|_{L^\infty(K)}+|D\phi|_{L^\infty(K)}\le c_K(1+M_1[\mu]). \tag{36}$$

Now, (22) and (36) imply that there exists a constant c_7 depending only on h, Λ, and $M_o[\mu]$ such that

$$\psi^-(\mathbf{z})\le R_o|\mathbf{z}|+c_7(1+M_1[\mu]) \tag{37}$$

for all $\mathbf{z}\in\mathbf{R}^d$. By (29) and (37) we have that there exists a constant c_8 depending only on h, Λ, and $M_o[\mu]$ such that

$$\int_{\mathbf{R}^d}|\psi(\mathbf{z})|d\mu(\mathbf{z})\le c_8(1+M_1[\mu]). \tag{38}$$

This concludes the proof of (ii).

Step 4. By (38),

$$\mu[B_R] \inf_{B_R} |\psi| \le c_8(1 + M_1[\mu]),$$

and so, if in addition $Lip(\psi) \le R_o$, we readily obtain (iii). This concludes the proof of the lemma. □

PROPOSITION 2.3. *Suppose that μ satisfies* (20) *such that (μ_n) is a sequence of Borel measures, that $M_o[\mu_n] = M_o[\mu]$ $(n = 1, 2, \ldots)$, that (μ_n) converges weak $*$ to μ, and that $(M_1[\mu_n])$ converges to $M_1[\mu]$. Then the following hold:*

(i) *There exists $(\psi_\mu, \phi_\mu) \in \mathcal{A}$ minimizing J_μ over $\mathcal{A}$, and*

$$\inf_{\mathcal{A}} J_\mu \le \liminf_{n \to +\infty} (\inf_{\mathcal{A}} J_{\mu_n}).$$

(ii) *We have that* $\limsup_{n \to +\infty} (\sup_{\Gamma(\mu_n)} \bar{I}) \le \sup_{\Gamma(\mu)} \bar{I}$.

(iii) *If $\sup_{\Gamma(\mu)} \bar{I} \ne -\infty$, then there exists $\gamma_\mu \in \Gamma(\mu)$ maximizing $\bar{I}$ over $\Gamma(\mu)$.*

Proof. Step 1. We shall show in Step 5 that (i) is a direct consequence of the following statement: If $(f_n, g_n) \in \mathcal{A}$ is such that $|\inf_{\mathcal{A}} J_{\mu_n} - J_{\mu_n}(f_n, g_n)| \le 1/n$, then there exists $(\psi_\mu, \phi_\mu) \in \mathcal{A}$ such that

$$J_\mu(\psi_\mu, \phi_\mu) \le \liminf_{n \to +\infty} J_{\mu_n}(f_n, g_n) \qquad (n = 1, 2, \ldots). \tag{39}$$

To proceed, let $R_1 > 0$ be such that

$$\mu[\mathrm{int}(B_{R_1})] > 1/2\mu[\mathbf{R}^d]. \tag{40}$$

Note that since (μ_n) converges weak $*$ to μ, in light of (40) we may assume without loss of generality that (see [13, p. 59])

$$\mu_n[\mathrm{int}(B_{R_1})] > 1/2M_o[\mu] = 1/2M_o[\mu_n] \tag{41}$$

for all $n = 1, 2, \ldots$. Define

$$\phi_n := (f_n)_\sharp, \qquad \psi_n := (\phi_n)^\sharp.$$

By Lemma A.1 (ii)–(iii) ψ_n and ϕ_n are convex functions, $\psi_n \le f_n$, $\phi_n \le g_n$, and

$$Lip(\psi_n) \le R_o; \tag{42}$$

hence

$$J_{\mu_n}(\psi_n, \phi_n) \le J_{\mu_n}(f_n, g_n) \tag{43}$$

for all $n = 1, 2, \ldots$. Since in addition $|\inf_{\mathcal{A}} J_{\mu_n} - J_{\mu_n}(\psi_n, \phi_n)| \le 1/n$, by Lemma 2.2 and (41) there exists a constant $\bar{c} > 0$ independent of n such that

$$\int_\Lambda |\phi_n(\mathbf{y})| d\mathbf{y} \le \bar{c} \tag{44}$$

and

$$|\psi_n(\mathbf{z})| \le R_o|\mathbf{z}| + \bar{c} \qquad (\mathbf{z} \in \mathbf{R}^d). \tag{45}$$

Using (45) we deduce that the sequence (ψ_n) is bounded in $W^{1,\infty}(B_{R'})$ for every $R' > 0$. Since ψ_n is convex, we may find a subsequence of (ψ_n) that we still label (ψ_n), converging in $L^\infty_{loc}(\mathbf{R}^d)$ to a convex function $\psi_\mu : \mathbf{R}^d \to \mathbf{R}$. One can readily check the following claims.

Step 2. Claim. We have that

$$\limsup_{n\to+\infty} \int_{B_R^c} (R_o|\mathbf{z}| + \bar{c}) d\mu_n(\mathbf{z}) \leq \int_{B_{R-2}^c} (R_o|\mathbf{z}| + \bar{c}) d\mu(\mathbf{z})$$

for all $R > 2$.

Step 3. Claim. We have that $\lim_{n\to+\infty} \int_{\mathbf{R}^d} |\psi_n - \psi_\mu| d\mu_n = 0$.

We next prove the following.

Step 4. Claim. We have that $\liminf_{n\to+\infty} \int_{\mathbf{R}^d} \psi_\mu d\mu_n \geq \int_{\mathbf{R}^d} \psi_\mu d\mu$.

Proof: For $R > 1$ let $l_R : \mathbf{R} \to [0,1]$ be of class C^∞ such that

$$l_R(t) = \begin{cases} 1 & \text{if} \quad |t| \leq R-1, \\ 0 & \text{if} \quad |t| \geq R. \end{cases} \tag{46}$$

We have that

$$\chi_{B_R^c} \leq 1 - l_R(|\mathbf{z}|) \leq \chi_{B_{R-2}^c}. \tag{47}$$

Because (μ_n) converges weak $*$ to μ and $(M_1[\mu_n])$ converges to $M_1[\mu]$, using (45) and (47) we have that

$$\begin{aligned} \liminf_{n\to+\infty} \int_{\mathbf{R}^d} \psi_\mu d\mu_n &\geq \int_{\mathbf{R}^d} \psi_\mu l_R d\mu - \int_{\mathbf{R}^d} (R_o|\mathbf{z}| + \bar{c})(1 - l_R(|\mathbf{z}|)) d\mu \\ &\geq \int_{\mathbf{R}^d} \psi_\mu l_R d\mu - \int_{B_{R-2}^c} (R_o|\mathbf{z}| + \bar{c}) d\mu. \end{aligned} \tag{48}$$

Letting R go to $+\infty$ in (48) we conclude the proof of Claim 4.

Now, combining Claims 3 and 4 we have that

$$\int_{\mathbf{R}^d} \psi_\mu d\mu \leq \liminf_{n\to+\infty} \int_{\mathbf{R}^d} \psi_n d\mu_n. \tag{49}$$

Similarly, since ϕ_n is convex (44) implies that there exists a convex function $\phi_\mu : conv(\Lambda) \to \mathbf{R} \cup \{+\infty\}$ such that up to a subsequence, (ϕ_n) converges pointwise to ϕ_μ in Λ and

$$\int_\Lambda \phi_\mu d\mathbf{y} \leq \liminf_{n\to+\infty} \int_\Lambda \phi_n d\mathbf{y}. \tag{50}$$

Because $(\psi_n, \phi_n) \in \mathcal{A}$, we obtain that $(\psi_\mu, \phi_\mu) \in \mathcal{A}$. Thanks to (43), (49), and (50) we have that

$$\inf_{\mathcal{A}} J_\mu \leq J_\mu(\psi_\mu, \phi_\mu) \leq \liminf_{n\to+\infty} J_{\mu_n}(f_n, g_n), \tag{51}$$

which proves (39).

Step 5. Taking $\mu_n \equiv \mu$ for all n in (51) we have that there exists $(\psi_\mu, \phi_\mu) \in \mathcal{A}$ minimizing J_μ over $\mathcal{A}$. Next, assuming (f_n, g_n) minimizes J_{μ_n} over $\mathcal{A}$, (51) implies that $\inf_{\mathcal{A}} J_\mu \leq \liminf_{n\to+\infty} (\inf_{\mathcal{A}} J_{\mu_n})$ which completes the proof of (i).

If $\limsup_{n\to+\infty}(\sup_{\Gamma(\mu_n)})\bar{I} = -\infty$, then (ii) is straightforward to obtain.

Step 6. Now we prove (ii). If $\limsup_{n\to+\infty}(\sup_{\Gamma(\mu_n)})\bar{I} = -\infty$, then (ii) is straightforward to obtain. Therefore we may assume without loss of generality that

$$\limsup_{n\to+\infty}(\sup_{\Gamma(\mu_n)}\bar{I}) > -\infty.$$

Note first that since by (21) $\sup_{\Gamma(\mu_n)}\bar{I} \leq \inf_{\mathcal{A}} J_{\mu_n}$, using the fact that $(M_1[\mu_n])$ converges to $M_1[\mu]$, and Lemma 2.2 (i) we have that $\limsup_{n\to+\infty}(\sup_{\Gamma(\mu_n)}\bar{I}) < +\infty$. Let (n_j) be such that

$$\limsup_{n\to+\infty}(\sup_{\Gamma(\mu_n)}\bar{I}) = \lim_{j\to+\infty}(\sup_{\Gamma(\mu_{n_j})}\bar{I}).$$

Choose e_1 a real number independent of j, smaller than $\sup_{\Gamma(\mu_{n_j})}\bar{I}$ for all $j \in \mathbf{N}$ and let $\gamma_{n_j} \in \Gamma(\mu_{n_j})$ be such that

$$\sup_{\Gamma(\mu_{n_j})}\bar{I} \leq \bar{I}[\gamma_{n_j}] + 1/n_j.$$

One can readily check that $\int_C h(\alpha)d\gamma_{n_j}$ is less than or equal to $R_o M_1[\mu_{n_j}] + 1 - e_1$, and so there exists a constant e_2 independent of j such that

$$\int_C |h(\alpha)|d\gamma_{n_j} \leq e_2 \tag{52}$$

for all $j \in \mathbf{N}$. By Proposition B.1, (52) implies that there exists a subsequence of (n_j) that we still label (n_j) and a Borel measure $\gamma \in \Gamma(\mu)$ such that (γ_{n_j}) converges weak $*$ to γ. Because h satisfies (4), $\bar{\Lambda}$ is contained in B_{R_o} and $\gamma_{n_j}[(0,+\infty)\times\Lambda^c\times\mathbf{R}^d] = 0$ we deduce that there exists a constant e_3 such that $m_R : (\alpha,\mathbf{y},\mathbf{z}) \to h(\alpha) - \mathbf{y}\cdot\mathbf{z} - e_3 + R_o|\mathbf{z}|$ is nonnegative for γ_{n_j}-almost every $(\alpha,\mathbf{y},\mathbf{z}) \in C$. Hence, if we define $k_R : (\alpha,\mathbf{y},\mathbf{z}) \to l_R(\alpha + |\mathbf{y}| + |\mathbf{z}|)$, then

$$\begin{aligned}\lim_{j\to+\infty}\int_C m_R d\gamma_{n_j} &\geq \lim_{j\to+\infty}\int_C m_R k_R d\gamma_{n_j}\\ &= \int_C m_R k_R d\gamma.\end{aligned} \tag{53}$$

Consequently,

$$\lim_{j\to+\infty}\int_C (h(\alpha) - \mathbf{y}\cdot\mathbf{z})d\gamma_{n_j} + R_o M_1[\mu_{n_j}] \geq \int_C (h(\alpha) - \mathbf{y}\cdot\mathbf{z})k_R d\gamma + R_o M_1[\mu]. \tag{54}$$

Letting R go to $+\infty$ in (54), using that $(M_1[\mu_{n_j}])$ converges to $M_1[\mu]$ we obtain that

$$\limsup_{n\to+\infty}(\sup_{\Gamma(\mu_n)}\bar{I}) \leq \bar{I}[\gamma] \leq \sup_{\Gamma(\mu)}\bar{I} \tag{55}$$

and conclude the proof of (ii).

Step 7. Setting $\mu_n = \mu$ for all $n \in \mathbf{N}$ in (55) we obtain (iii). □

THEOREM 2.1 (duality). *Suppose that h satisfies* (3), (4), (5) *and that μ satisfies* (20). *Then the following hold:*

(i) *There exists a pair (ψ_μ, ϕ_μ) of convex functions minimizing J_μ over $\mathcal{A}$ such that $(\psi_\mu)_\sharp = \phi_\mu$ and $(\phi_\mu)^\sharp = \psi_\mu$ and $Lip(\psi_\mu) \le R_o$.*

(ii) *The duality relation $\sup_{\Gamma(\mu)} \bar{I} = \inf_{\mathcal{A}} J_\mu$ holds. Defining on C the measure γ by*

$$\int_C g d\gamma = \int_\Lambda \frac{1}{\beta_\mu(\mathbf{y})} g(\beta_\mu(\mathbf{y}), \mathbf{y}, D\psi_\mu^*(\mathbf{y})) d\mathbf{y}$$

for all $g \in C_o(\mathbf{R} \times \mathbf{R}^d \times \mathbf{R}^d)$, we have that γ is the unique maximizer of $\bar{I}$ over $\Gamma(\mu)$. Here $\beta_\mu : \Lambda \to (0, +\infty)$ is a Borel map such that $\beta_\mu(\mathbf{y})(\psi_\mu)_\sharp(\mathbf{y}) + \psi_\mu(D\psi_\mu^*(\mathbf{y})) = \mathbf{y} \cdot D\psi_\mu^*(\mathbf{y}) - h(\beta_\mu(\mathbf{y}))$ for almost every $\mathbf{y} \in \Lambda$.

(iii) *If we assume in addition that $\mu[N] = 0$ for every $(d-1)$-rectifiable subset N of $\mathbf{R}^d$, then γ is of the form $\gamma = \gamma^{D\psi}$, i.e., γ can be parametrized on $(\mathbf{R}^d, \mu)$:*

$$\int_C g d\gamma = \int_{\mathbf{R}^d} g(\beta_\mu(D\psi_\mu(\mathbf{z})), D\psi_\mu(\mathbf{z}), \mathbf{z}) d\mu(\mathbf{z})$$

for all $g \in C_o(\mathbf{R} \times \mathbf{R}^d \times \mathbf{R}^d)$.

Proof. By Proposition 2.3 there exists a pair (ψ_μ, ϕ_μ) minimizing J over $\mathcal{A}$. By Lemma A.1 (iii)–(iv) the pairs $(\psi_\mu, (\psi_\mu)_\sharp)$ and $(((\psi_\mu)_\sharp)^\sharp, (\psi_\mu)_\sharp)$ minimize J over $\mathcal{A}$ and $(((\psi_\mu)_\sharp)^\sharp)_\sharp = (\psi_\mu)_\sharp$. Hence, we may assume without loss of generality that ψ_μ, ϕ_μ are convex, $(\psi_\mu)_\sharp = \phi_\mu$, and $(\phi_\mu)^\sharp = \psi_\mu$, and so

$$Lip(\psi_\mu) \le R_o \tag{56}$$

(see Lemma A.1). This concludes the proof of (i).

Step 1. We first give the proof of (ii) in the special case when there exists $R > 0$ such that the support of μ is contained in B_R and $\mu[N] = 0$ for every $(d-1)$-rectifiable subset N of $\mathbf{R}^d$.

Step 2. For $G \in C_o(\mathbf{R}^d)$ and $r > 0$ define

$$\psi_r(\mathbf{z}) := \begin{cases} \psi_\mu(\mathbf{z}) + rG(\mathbf{z}) & \text{if} \quad \mathbf{z} \in B_R, \\ +\infty & \text{if} \quad \mathbf{z} \notin B_R \end{cases}$$

and

$$\phi_r := (\psi_r)_\sharp.$$

We have that ψ_r^* is finite at every point of $\mathbf{R}^d$ and so $D\psi_r^*$ exists except on a $(d-1)$-rectifiable set (see [1]). Hence, $S_r := D\psi_r^* : \Lambda \to B_R$ is well-defined μ-almost everywhere. In light of Lemma A.1 let $\beta_r : \Lambda \to (0, +\infty)$ be the unique Borel function such that

$$\beta_r(\mathbf{y})\phi_r(\mathbf{y}) + \psi_r(S_r(\mathbf{y})) = \mathbf{y} \cdot S_r(\mathbf{y}) - h(\beta_r(\mathbf{y})). \tag{57}$$

Note that β_r is well-defined μ-almost everywhere. By (56) $|\psi_r|_{L^\infty(B_R)}$ is bounded independently of $|r| \le 1$ and so Lemma A.1 implies

$$c \le \beta_r(\mathbf{y}) \le 1/c \tag{58}$$

for all $\mathbf{y} \in \Lambda$ and for some constant $c > 0$ independent of r. Observe that (57) implies

$$-\frac{r}{\beta_o(\mathbf{y})} G(S_o(\mathbf{y})) \leq \phi_r(\mathbf{y}) - \phi_o(\mathbf{y}) \leq -\frac{r}{\beta_r(\mathbf{y})} G(S_r(\mathbf{y})) \tag{59}$$

for all $\mathbf{y} \in \Lambda$. This, together with (58), yields

$$|\phi_r(\mathbf{y}) - \phi_o(\mathbf{y})| \leq \frac{r}{c} |G|_{L^\infty(\mathbf{R}^d)} \tag{60}$$

for all $\mathbf{y} \in \Lambda$.

Step 3. Claim. Whenever $S_o(\mathbf{y})$ exists we have that $(\phi_r(\mathbf{y}) - \phi_o(\mathbf{y}))/r$ tends to $-G(S_o(\mathbf{y}))/\beta_o(\mathbf{y})$ as r tends to 0.

Proof. Fix $\mathbf{y}$ such that $S_o(\mathbf{y})$ exists and assume that $(r_j) \subset (0, +\infty)$ is a sequence converging to 0,

$$S_{r_j}(\mathbf{y}) \to \mathbf{z}_o, \qquad \beta_{r_j}(\mathbf{y}) \to \alpha_o, \tag{61}$$

as j tends to $+\infty$. Since (ψ_r) converges uniformly to ψ_o on B_R and by (60) (ϕ_r) converges uniformly to ϕ_o on Λ, (57) implies that

$$\alpha_o \phi_o(\mathbf{y}) + \psi_o(\mathbf{z}_o) = \mathbf{y} \cdot \mathbf{z}_o - h(\alpha_o). \tag{62}$$

Since $S_o(\mathbf{y}) = D\psi_o^*(\mathbf{y})$ exists, (62) and Lemma A.1 imply

$$\alpha_o = \beta_o(\mathbf{y}) \quad \text{and} \quad \mathbf{z}_o = S_o(\mathbf{y}).$$

Because $(r_j) \subset (0, +\infty)$ is arbitrary we deduce that $(S_r(\mathbf{y}))$ converges to $S_o(\mathbf{y})$ and $(\beta_r(\mathbf{y}))$ converges to $\beta_o(\mathbf{y})$ as r tends to 0. This together with (59) yields Claim 3.

Step 4. Claim. S_o pushes $d\mathbf{y}/\beta_o(\mathbf{y})$ forward to μ.

Proof. Note that $J_\mu[\psi_o, \phi_o] = J_\mu[\psi_\mu, \phi_\mu]$ and so (ψ_o, ϕ_o) also minimizes J_μ over $\mathcal{A}$. This combined with Claim 3 implies

$$0 = \lim_{r \to 0} \frac{J_\mu[\psi_r, \phi_r] - J_\mu[\psi_o, \phi_o]}{r} = \int_{\mathbf{R}^d} G d\mu - \int_\Lambda \frac{G \circ S_o}{\beta_o} d\mathbf{y}. \tag{63}$$

Since G is arbitrary in (63), we conclude Claim 4.

Step 5. Using (57) and Claim 4 we have that

$$\begin{aligned} J_\mu[\psi_o, \phi_o] = \int_\Lambda \frac{\psi_o \circ S_o + \beta_o \phi_o}{\beta_o} d\mathbf{y} &= \int_\Lambda \frac{\mathbf{y} \cdot S_o(\mathbf{y}) - h(\beta_o(\mathbf{y}))}{\beta_o(\mathbf{y})} d\mathbf{y} \\ &= \int_C (\mathbf{y} \cdot \mathbf{z} - h(\alpha)) d\gamma_\mu = \bar{I}[\gamma_\mu], \end{aligned} \tag{64}$$

where we have defined the measure γ_μ by

$$\int_C g d\gamma_\mu = \int_\Lambda \frac{1}{\beta_o(\mathbf{y})} g(\beta_o(\mathbf{y}), \mathbf{y}, S_o(\mathbf{y})) d\mathbf{y}$$

for all $g \in C_o(\mathbf{R} \times \mathbf{R}^d \times \mathbf{R}^d)$. Clearly $\gamma_\mu \in \Gamma(\mu)$. Combining (21) and (64) we deduce that

$$\sup_{\Gamma(\mu)} \bar{I} = \inf_{\mathcal{A}} J_\mu. \tag{65}$$

Step 6. We complete the proof of (ii). Assume now that μ satisfies only (20). Let (μ_n) be a sequence of Borel measures on $\mathbf{R}^d$ such that $\mu_n[N] = 0$ whenever N is a $(d-1)$-rectifiable subset of $\mathbf{R}^d$, $M_o[\mu_n] = M_o[\mu]$, $spt(\mu_n)$ is bounded for all $n = 1, 2, \ldots$, and $(M_1[\mu_n])$ converges to $M_1[\mu]$ as n tends to $+\infty$. Combining Proposition 2.3 and (65) we have that

$$\inf_{\mathcal{A}} J_\mu \le \liminf_{n\to+\infty}(\inf_{\mathcal{A}} J_{\mu_n}) \le \limsup_{n\to+\infty}(\sup_{\Gamma(\mu_n)} \bar{I}) \le \sup_{\Gamma(\mu)} \bar{I}. \tag{66}$$

Combining (21) and (66) we deduce that

$$\sup_{\Gamma(\mu)} \bar{I} = \inf_{\mathcal{A}} J_\mu.$$

This proves that duality persists under the sole assumption that μ satisfies only (20). In light of Proposition 2.3 and the above duality result, if γ maximizes $\bar{I}$ over $\Gamma(\mu)$, we have that

$$\int_C (\psi_\mu(\mathbf{z}) + \alpha\phi_\mu(\mathbf{y}) + h(\alpha) - \mathbf{y}\cdot\mathbf{z}) d\gamma = 0,$$

and so

$$\psi_\mu(\mathbf{z}) + \alpha\phi_\mu(\mathbf{y}) + h(\alpha) - \mathbf{y}\cdot\mathbf{z} = 0$$

for every $(\alpha, \mathbf{y}, \mathbf{z}) \in D'$ where $D' \subset C$ is such that $\gamma[C \setminus D'] = 0$. Let A be the subset of Λ where $D\psi_\mu^*$ exists. Since $H^d[\Lambda \setminus A] = 0$ we deduce that $\gamma[C \setminus D'' = 0$ where

$$D'' := (0, +\infty) \times A \times \mathbf{R}^d.$$

In light of Lemma A.1, there exists a Borel function $\beta_\mu : \Lambda \to (0, +\infty)$ such that

$$D := D' \cap D'' \subset \{(\beta_\mu(\mathbf{y}), \mathbf{y}, D\psi_\mu^*(\mathbf{y})) \mid \mathbf{y} \in A\}. \tag{67}$$

Since $\gamma[C \setminus D] = 0$, (67) implies the representation formula

$$\int_C g d\gamma = \int_\Lambda \frac{1}{\beta_\mu(\mathbf{y})} g(\beta_\mu(\mathbf{y}), \mathbf{y}, D\psi_\mu^*(\mathbf{y})) d\mathbf{y} \tag{68}$$

for all $g \in C_o(\mathbf{R} \times \mathbf{R}^d \times \mathbf{R}^d)$, and so γ is uniquely determined. This concludes the proof of (ii).

Step 7. We complete the proof of (iii). Assume that μ satisfies (20) and $\mu[N] = 0$ whenever N is a $(d-1)$-rectifiable subset of $\mathbf{R}^d$. Since $\gamma[C]$ is finite, (68) implies that $1/\beta_\mu \in L^1(\Lambda)$. Choosing $g \equiv g(\mathbf{z})$ in (68) we obtain that $D\psi_\mu^*$ is the optimal map in the Monge problem that pushes $d\mathbf{y}/\beta_\mu(\mathbf{y})$ forward to μ, and so $D\psi_\mu^*$ is one-to-one with respect to Lebesgue measure, its inverse is $D\psi_\mu$ and is one-to-one with respect to μ (see Proposition D.1). This together with the representation formula of γ given in (ii) proves (iii). □

Remark 2.4. Note that if h satisfies (3), (4), (5) and μ is a measure whose support is contained in B_R for some $R > 0$, then by Step 1 of the proof of Theorem 2.1 we obtain that ψ_μ^* can be extended to a convex, lower semicontinuous function which is finite on $\mathbf{R}^d$. If $\beta_\mu : \Lambda \to (0, +\infty)$ is the Borel function such that $\beta_\mu(\mathbf{y})(\psi_\mu)_\sharp(\mathbf{y}) + \psi_\mu(D\psi_\mu^*(\mathbf{y})) = \mathbf{y}\cdot D\psi_\mu^*(\mathbf{y}) - h(\beta_\mu(\mathbf{y}))$ for almost every $\mathbf{y} \in \Lambda$, and ψ_μ is convex, lower semicontinuous, since $H \circ \beta_\mu = \psi_\mu^*$, we then deduce that there exists a constant $c > 0$ such that $c \le \beta_\mu \le 1/c$.

3. Existence of equilibrium configuration. Throughout this section we assume that $\Omega, \Lambda \subset \mathbf{R}^d$ are two open bounded sets whose closures are contained in the closed ball B_{R_o} of center 0 and radius R_o. We assume that h satisfies (3), (4), (5) and $\mathbf{F} \in L^1(\Omega)^d$ is a Borel map. The aim of this section is to prove that a direct consequence of section 2 is that problem

$$\inf_{(\psi,\phi)\in\mathcal{A}} J[\psi,\phi] \tag{69}$$

and problem

$$\sup_{\mathbf{u}\in\mathcal{U}'_\Lambda} I[\mathbf{u}] \tag{70}$$

are dual of each other. Here

$$I[\mathbf{u}] := \int_\Omega (\mathbf{F}\cdot\mathbf{u} - h(|\det D\mathbf{u}|))d\mathbf{x} \qquad (\mathbf{u}\in\mathcal{U}'_\Lambda),$$

and J is defined as in (14) by

$$J[\psi,\phi] := \int_\Omega \psi(\mathbf{F}(\mathbf{x}))d\mathbf{x} + \int_\Lambda \phi(\mathbf{y})d\mathbf{y}.$$

We also show that if in addition $\mathbf{F}$ is one-to-one almost everywhere and $|\mathbf{F}^{-1}(N)| = 0$ whenever N is $(d-1)$-rectifiable, then (70) admits a unique minimizer. The inequality

$$\sup_{\mathbf{u}\in\mathcal{U}'_\Lambda} I[\mathbf{u}] \le \inf_{(\psi,\phi)\in\mathcal{A}} J[\psi,\phi]$$

is straightforward. Indeed, if $\mathbf{u}\in\mathcal{U}'_\Lambda$ and $(\psi,\phi)\in\mathcal{A}$, then

$$\mathbf{F}\cdot\mathbf{u} - h(|\det D\mathbf{u}|) \le \psi\circ\mathbf{F} + |\det D\mathbf{u}|\cdot\phi\circ\mathbf{u}$$

almost everywhere in Ω, which by integration yields $I[\mathbf{u}] \le J[\psi,\phi]$. Because $\mathbf{u}\in\mathcal{U}'_\Lambda$ and $(\psi,\phi)\in\mathcal{A}$ are arbitrary we have that

$$\sup_{\mathbf{u}\in\mathcal{U}'_\Lambda} I[\mathbf{u}] \le \inf_{(\psi,\phi)\in\mathcal{A}} J[\psi,\phi]. \tag{71}$$

The task in this section is to establish the reverse inequality.

LEMMA 3.1. *Suppose that* (3), (4), *and* (5) *hold and that* $\psi_o : \mathbf{R}^d \to \mathbf{R}$ *is convex, lower semicontinuous. If* $\bar{\mathbf{u}}\in\mathcal{U}'_\Lambda$, $\mathbf{F} = D\psi_o^*\circ\bar{\mathbf{u}}$, *and* $H(|\det D\bar{\mathbf{u}}|) = (\psi_o)^*\circ\bar{\mathbf{u}}$, *then* $I[\bar{\mathbf{u}}] = J[\psi_o,(\psi_o)_\sharp]$, $\bar{\mathbf{u}}$ *is a maximizer of* I *over* $\mathcal{U}'_\Lambda$, *and the pair* $(\psi_o,(\psi_o)_\sharp)$ *minimizes* J *over* $\mathcal{A}$.

Proof. Define $\phi_o := (\psi_o)_\sharp$. Because $|\det D\bar{\mathbf{u}}| \neq 0$ almost everwhere in the weak sense, we have that $|\bar{\mathbf{u}}^{-1}[N]| = 0$ whenever $|N| = 0$. Also, since the convex functions ϕ_o and $(\psi_o)^*$ are differentiable everywhere except on a $(d-1)$-rectifiable set, we have that both ϕ_o and $(\psi_o)^*$ are differentiable at $\bar{\mathbf{u}}(\mathbf{x})$ for almost every $\mathbf{x}\in\Omega$. By Lemma A.1, for these $\mathbf{x}\in\Omega$ we may define $\alpha(\mathbf{x}) > 0$ and $\mathbf{z}(\mathbf{x}) \in \partial\psi_o^*(\bar{\mathbf{u}}(\mathbf{x}))$ such that

$$H(\alpha(\mathbf{x})) = \psi_o^*(\bar{\mathbf{u}}(\mathbf{x})) \tag{72}$$

and

$$\alpha(\mathbf{x})\phi_o(\bar{\mathbf{u}}(\mathbf{x})) + \psi_o(\mathbf{z}(\mathbf{x})) = \mathbf{z}(\mathbf{x})\cdot\bar{\mathbf{u}}(\mathbf{x}) - h(\alpha(\mathbf{x})). \tag{73}$$

We use the fact that H is decreasing, $H(|\det D\bar{\mathbf{u}}|) = \psi_o^* \circ \bar{\mathbf{u}}$, and (72) to obtain that

$$\alpha(\mathbf{x}) = |\det D\bar{\mathbf{u}}(\mathbf{x})|. \tag{74}$$

Since ψ_o^* is differentiable at $\bar{\mathbf{u}}(\mathbf{x})$ and $\mathbf{z}(\mathbf{x}) \in \partial\psi_o^*(\bar{\mathbf{u}}(\mathbf{x}))$ we deduce that

$$\mathbf{z}(\mathbf{x}) = \mathbf{F}(\mathbf{x}). \tag{75}$$

By (73), (74), and (75) we obtain that

$$|\det D\bar{\mathbf{u}}(\mathbf{x})|\phi_o(\bar{\mathbf{u}}(\mathbf{x})) + \psi_o(\mathbf{F}(\mathbf{x})) = \mathbf{F}(\mathbf{x}) \cdot \bar{\mathbf{u}}(\mathbf{x}) - h(|\det D\bar{\mathbf{u}}(\mathbf{x})|),$$

which by integration yields $I[\bar{\mathbf{u}}] = J[\psi_o, \phi_o]$. Since $(\psi_o, \phi_o) \in \mathcal{A}$ (71) implies $\bar{\mathbf{u}}$ maximizes I over $\mathcal{U}'_\Lambda$ and $(\psi_o, (\psi_o)_\sharp)$ minimizes J over $\mathcal{A}$. □

THEOREM 3.1 (main results). *Suppose that* (3), (4), *and* (5) *hold. Then we have the following.*

(i) $\inf_{\mathcal{A}} J[\psi, \phi] = \sup_{\mathcal{U}'_\Lambda} I[\mathbf{u}]$.

(ii) *If* $\mathbf{F}$ *is one-to-one almost everywhere and* $(d-1)$*-nondegenerate, then* I *admits a unique maximizer* $\bar{\mathbf{u}}$ *over* $\mathcal{U}'_\Lambda$, $\bar{\mathbf{u}} = D\psi_\mu \circ \mathbf{F}$, *and* $I[\bar{\mathbf{u}}] = J[\psi_\mu, (\psi_\mu)_\sharp]$, *and the map* $\bar{\mathbf{u}}$ *satisfies the Hamilton–Jacobi equation* $H(|\det D\bar{\mathbf{u}}|) = \psi_\mu^* \circ \bar{\mathbf{u}}$ *for some lower semicontinuous convex function* $\psi_\mu : \mathbf{R}^d \to \mathbf{R}$ *such that* $Lip(\psi_\mu) \le R_o$ *and* $\psi_\mu = ((\psi_\mu)_\sharp)^\sharp$.

(iii) *If* $\mathbf{F}$ *satisfies the assumptions in* (ii) *and in addition* $\mathbf{F} \in L^\infty(\Omega)^d$, *then there exists a constant* $c > 0$ *such that* $c \le |\det D\bar{\mathbf{u}}| \le 1/c$, *and we may extend* ψ_μ^* *into a Lipschitz, convex function in a neighborhood of* $conv(\bar{\Lambda})$.

Proof. We define on $\mathbf{R}^d$ the measure μ given by

$$\mu[A] := |\mathbf{F}^{-1}[A]|$$

for $A \subset \mathbf{R}^d$. Note that

$$J[\psi, \phi] = \int_{\mathbf{R}^d} \psi d\mu + \int_\Lambda \phi d\mathbf{y},$$

which, using the notation of section 2, is $J_\mu[\psi, \phi]$, and the following condition on the moments is satisfied:

$$M_o[\mu] = |\Omega| < +\infty, \qquad M_1[\mu] = |\mathbf{F}|_{L^1(\Omega)} < +\infty. \tag{76}$$

By Theorem 2.1 (i) there exists a pair (ψ_μ, ϕ_μ) of convex functions minimizing J_μ over $\mathcal{A}$ such that $(\psi_\mu)_\sharp = \phi_\mu$ and $(\phi_\mu)^\sharp = \psi_\mu$ and $Lip(\psi_\mu) \le R_o$.

Step 1. Assume first that $\mathbf{F}$ is one-to-one almost everywhere, $(d-1)$-nondegenerate. Note that $\mu[N] = 0$ whenever N is a $(d-1)$-rectifiable subset of $\mathbf{R}^d$. Since ψ_μ is convex, the set where ψ_μ is not differentiable is $(d-1)$-rectifiable (see [1]) and so

$$\bar{\mathbf{u}}(\mathbf{x}) := D\psi_\mu(\mathbf{F}(\mathbf{x})) \tag{77}$$

is defined for almost every $\mathbf{x} \in \Omega$. In light of Theorem 2.1 (iii) $D\psi_\mu$ is the optimal map in the Monge problem that pushes μ forward to $d\mathbf{y}/\beta_\mu(\mathbf{y})$ where $\beta_\mu : \Lambda \to (0, +\infty)$ is a Borel function such that

$$\beta_\mu(\mathbf{y})(\psi_\mu)_\sharp(\mathbf{y}) + \psi_\mu(D\psi_\mu^*(\mathbf{y})) = \mathbf{y} \cdot D\psi_\mu^*(\mathbf{y}) - h(\beta_\mu(\mathbf{y}))$$

for almost every $\mathbf{y} \in \Lambda$. Note that in light of Remark 2.4, if in addition $\mathbf{F} \in L^\infty(\Omega)^d$, then we may assume without loss of generality that ψ_μ^* is Lipschitz on $conv(\bar{\Lambda})$. We have that $D\psi_\mu$ is one-to-one on $\mathbf{R}^d$ up to a set of zero measure with respect to μ and $D\psi_\mu$ maps $\mathbf{R}^d$ onto Λ. We deduce that

(78) $\quad \bar{\mathbf{u}}$ is one-to-one up to a set of zero measure with respect to $\chi_\Omega d\mathbf{x}$.

Recall that in light of Theorem 2.1 (iii) the measure γ_μ defined on C by

$$\int_C g d\gamma_\mu = \int_{\mathbf{R}^d} g(\beta_\mu(D\psi_\mu(\mathbf{z})), D\psi_\mu(\mathbf{z}), \mathbf{z}) d\mu(\mathbf{z})$$

for all $g \in C_o(\mathbf{R} \times \mathbf{R}^d \times \mathbf{R}^d)$ maximizes $\bar{I}$ over $\Gamma(\mu)$. Therefore we have that

$$\int_\Omega f(\bar{\mathbf{u}}(\mathbf{x}))\beta_\mu(\bar{\mathbf{u}}(\mathbf{x}))d\mathbf{x} = \int_\Lambda f(\mathbf{y})d\mathbf{y}$$

for all $f \in C_o(\mathbf{R}^d)$. Consequently,

$$|\det D\bar{\mathbf{u}}| = \beta_\mu \circ \bar{\mathbf{u}}. \tag{79}$$

Using (78), (79), and the fact that $\beta_\mu > 0$ we obtain that

$$\bar{\mathbf{u}} \in \mathcal{U}'_\Lambda. \tag{80}$$

Since $\beta_\mu(\psi_\mu)_\sharp + \psi_\mu \circ D\psi_\mu^* = \mathbf{id} \cdot D\psi_\mu^* - h \circ \beta_\mu$ Lemma A.1 implies $H \circ \beta_\mu = \psi_\mu^*$, and so using (79) we obtain that

$$H(|\det D\bar{\mathbf{u}}|) = \psi_\mu^* \circ \bar{\mathbf{u}}. \tag{81}$$

By Lemma 3.1, (77), (80), and (81) we obtain that $\bar{\mathbf{u}}$ maximizes I over $\mathcal{U}'_\Lambda$ and $I[\bar{\mathbf{u}}] = J[\psi_\mu, (\psi_\mu)_\sharp]$. Therefore, we have proved (i) under the assumption that $\mathbf{F}$ is one-to-one almost everywhere, $(d-1)$-nondegenerate.

Step 2. We prove that $\bar{\mathbf{u}}$ is the unique maximizer of I over $\mathcal{U}'_\Lambda$. Indeed, if $\mathbf{u}$ is another maximizer of I over $\mathcal{U}'_\Lambda$, the duality relation between (10) and (13) implies

$$\mathbf{F}(\mathbf{x}) \cdot \mathbf{u}(\mathbf{x}) - h(|\det D\mathbf{u}(\mathbf{x})|) = \psi_\mu(\mathbf{F}(\mathbf{x})) + |\det D\mathbf{u}(\mathbf{x})|\phi_\mu(\mathbf{u}(\mathbf{x}))$$

for all almost every $\mathbf{x} \in \Omega$, and so, by Lemma A.1 (i),

$$\mathbf{u}(\mathbf{x}) \in \partial\psi_\mu(\mathbf{F}(\mathbf{x})) \tag{82}$$

for these $\mathbf{x}$. Since ψ_μ is differentiable everywhere in B_R except on a $(d-1)$-rectifiable set and $\mathbf{F}$ is $(d-1)$-nondegenerate, (82) implies $\mathbf{u}(\mathbf{x}) = D\psi_\mu(\mathbf{F}(\mathbf{x})) = \bar{\mathbf{u}}(\mathbf{x})$ for all almost every $\mathbf{x} \in \Omega$. This concludes the proof of (ii).

Step 3. If $\mathbf{F}$ satisfies the assumptions in (ii) and in addition $\mathbf{F} \in L^\infty(\Omega)^d$, then there exists $R > 0$ such that the support of μ is contained in B_R. Using Remark 2.4 and (79) we obtain (iii).

Step 4. We now prove (i) under the sole assumption that $\mathbf{F} \in L^1(\Omega)^d$. For each $n \in \mathbf{N}$ we may find $\mathbf{F}_n \in L^\infty(\Omega)^d$ that is one-to-one almost everywhere, $(d-1)$-nondegenerate, and such that

$$|\mathbf{F}_n - \mathbf{F}|_{L^1(\Omega)} \to 0$$

as n tends to $+\infty$. Define

$$J_n[\psi,\phi] := \int_{\mathbf{R}^d} \psi(\mathbf{F}_n(\mathbf{x}))d\mathbf{x} + \int_\Lambda \phi(\mathbf{y})d\mathbf{y}$$

and

$$I_n[\mathbf{u}] := \int_\Omega (\mathbf{F}_n \cdot \mathbf{u} - h(|\det D\mathbf{u}|))d\mathbf{x}.$$

By (ii) there exists $\psi_n : \mathbf{R}^d \to \mathbf{R}$ convex function such that $Lip(\psi_n) \le R_o$ and

$$J_n[\psi_n, (\psi_n)_\sharp] = \inf_{\mathcal{A}} J_n = \sup_{\mathcal{U}'_\Lambda} I_n. \tag{83}$$

Using that $Lip(\psi_n) \le R_o$ we have that

$$J_n[\psi_n, (\psi_n)_\sharp] \ge \inf_{\mathcal{A}} J - R_o|\mathbf{F}_n - \mathbf{F}|_{L^1(\Omega)} \tag{84}$$

and using that $\mathbf{u}(\Omega) \subset \Lambda \subset B_{R_o}$ for all $\mathbf{u} \in \mathcal{U}'_\Lambda$ we deduce that

$$\sup_{\mathcal{U}'_\Lambda} I_n \le \sup_{\mathcal{U}'_\Lambda} I + R_o|\mathbf{F}_n - \mathbf{F}|_{L^1(\Omega)}. \tag{85}$$

Combining (83), (84), and (85) we obtain (i). □

COROLLARY 3.2 (characterization of maximizers of I). *Suppose that* (3), (4), (5) *hold and that* $\mathbf{F} \in L^1(\Omega)^d$. *Assume that* $\bar{\mathbf{u}} \in \mathcal{U}'_\Lambda$. *Then* $\bar{\mathbf{u}}$ *maximizes* I *over* $\mathcal{U}'_\Lambda$ *if and only if there exists a lower semicontinuous convex function* $\psi_o : \mathbf{R}^d \to \mathbf{R}$ *such that* $D\psi_o^*$ *exists almost everywhere in* Λ, $\mathbf{F} = D\psi_o^* \circ \bar{\mathbf{u}}$, *and* $H(|\det D\bar{\mathbf{u}}|) = \psi_o^* \circ \bar{\mathbf{u}}$ *on* Ω.

Proof. Step 1. Assume that $\bar{\mathbf{u}}$ maximizes I over $\mathcal{U}'_\Lambda$. By Theorem 3.1 there exists a lower semicontinuous convex function $\psi_o : \mathbf{R}^d \to \mathbf{R}$ such that $I[\bar{\mathbf{u}}] = J[\psi_o, \phi_o]$ and $\psi_o = (\phi_o)^\sharp$, where $\phi_o := (\psi_o)_\sharp$. We deduce that

$$|\det D\bar{\mathbf{u}}(\mathbf{x})|\phi_o(\bar{\mathbf{u}}(\mathbf{x})) + \psi_o(\mathbf{F}(\mathbf{x})) = \mathbf{F}(\mathbf{x}) \cdot \bar{\mathbf{u}}(\mathbf{x}) - h(|\det D\bar{\mathbf{u}}(\mathbf{x})|)) \tag{86}$$

for almost every $\mathbf{x} \in \Omega$. Since ψ_o^* is differentiable at almost every $\bar{\mathbf{u}}(\mathbf{x})$, using (86) and Lemma A.1 we deduce that

$$\mathbf{F} = D\psi_o^* \circ \bar{\mathbf{u}} \tag{87}$$

and

$$H(|\det D\bar{\mathbf{u}}|) = \psi_o^* \circ \bar{\mathbf{u}}. \tag{88}$$

Step 2. The converse implication is given by Lemma 3.1, and we conclude the proof of the lemma. □

4. Smoothness of equilibrium configurations. Throughout this section, unless the contrary is explicitly stated, we assume that $\Omega, \Lambda \subset \mathbf{R}^d$ are two open bounded sets. Recall that $d \ge 2$ is an integer. We now state the main result of this section.

THEOREM 4.1 (smoothness of maximizers of I). *Assume that* Ω *is connected, its boundary* $\partial\Omega$ *is Lipschitz,* Λ *and* $\mathbf{F}(\bar{\Omega})$ *are convex. Assume that* $\mathbf{F}, \det D\mathbf{F} \in C^1(\bar{\Omega})^d$, $0 < \det D\mathbf{F}$ *on* $\bar{\Omega}$, $\mathbf{F}$ *is a homeomorphism of* $\bar{\Omega}$ *onto* $\mathbf{F}(\bar{\Omega})$. *If* h *satisfies* (3), (4), *and* (5), *then the following hold:*

(i) *Problem* $\sup_{\mathcal{A}} -J$ *and* $\inf_{\mathcal{U}_\Lambda} E$ *are dual to each other, and there exists a unique* $\bar{\mathbf{u}}$ *minimizing* E *over* $\mathcal{U}_\Lambda$.

(ii) *We have that* $\bar{\mathbf{u}} \in C^1(\Omega)^d \cap C^{0,s}(\bar{\Omega})^d$, $\det D\bar{\mathbf{u}} \in C^{0,s}(\bar{\Omega}) \cap C^1(\Omega)$ *for all* $0 < s < 1$, *and* $\det D\bar{\mathbf{u}} + 1/\det D\bar{\mathbf{u}} \in L^\infty(\Omega)$.

(iii) *Furthermore,* $\bar{\mathbf{u}}$ *satisfies the partial differential equations* (9) *in the weak sense and* (11) *pointwise.*

Proof. Step 1. To show (i), it suffices to check that the map $\bar{\mathbf{u}}$ maximizing I over $\mathcal{U}'_\Lambda$ belongs to $\mathcal{U}_\Lambda$. By Theorem 3.1 there exists a lower semicontinuous, convex function $\psi_o : \mathbf{R}^d \to \mathbf{R}$ such that $\bar{\mathbf{u}} := D\psi_o \circ \mathbf{F} \in \mathcal{U}'_\Lambda$, $H \circ |\det D\bar{\mathbf{u}}| = \psi_o^* \circ \bar{\mathbf{u}}$,

$$I[\bar{\mathbf{u}}] = J[\psi_o, (\psi_o)_\sharp], \tag{89}$$

and $D\psi_o$ pushes $f_o d\mathbf{z}$ forward to $d\mathbf{y}/\beta_o(\mathbf{y})$, where

$$f_o(\mathbf{z}) := \frac{1}{\det D\mathbf{F}(\mathbf{F}^{-1}(\mathbf{z}))} \qquad (\mathbf{z} \in \mathbf{F}(\bar{\Omega})),$$

and $\beta_o : \Lambda \to (0, +\infty)$ is defined by $\beta_o(\mathbf{y})(\psi_o)_\sharp(\mathbf{y}) + \psi_o(D\psi_o^*(\mathbf{y})) = \mathbf{y} \cdot D\psi_o^*(\mathbf{y}) - h(\beta_o(\mathbf{y}))$. By Lemma A.1 we have that

$$H \circ \beta_o = \psi_o^*. \tag{90}$$

Since $\mathbf{F}$ is bounded we may assume without loss of generality that ψ_o^* is Lipschitz on $\bar{\Lambda}$ and because the inverse H^{-1} of H is of class C^1, (90) and Proposition D.2 imply that $\beta_o \in C^1(\bar{\Lambda})$. Clearly, f_o is of class C^1, bounded below and above on $\mathbf{F}(\bar{\Omega})$. Using Proposition D.2 again, using that $D\psi_o$ pushes $f_o d\mathbf{z}$ forward to $d\mathbf{y}/\beta_o(\mathbf{y})$ and that the density functions f_o and $1/\beta_o(\mathbf{y})$ are smooth we deduce that $D\psi_o \in C^{0,s}(\mathbf{F}(\bar{\Omega}))^d \cap C^{1,s}(\mathbf{F}(\Omega))^d$ for all $0 < s < 1$. This proves (i) and (ii). Note that there exists a constant $c > 0$ such that

$$c \le \det D\bar{\mathbf{u}} \le 1/c. \tag{91}$$

Step 2. Let $\mathbf{v} \in C_o^\infty(\Omega)^d$, let K be the support of $\mathbf{v}$, and for each $|r| < 1$ define

$$\mathbf{u}_r := \bar{\mathbf{u}} + r\mathbf{v}.$$

Since $\bar{\mathbf{u}} \in C^1(K)$, $\mathbf{u}_r = \bar{\mathbf{u}}$ on $\Omega \setminus K$, and (91) holds we deduce that $(D\mathbf{u}_r)$ converges uniformly to $D\bar{\mathbf{u}}$ on Ω and there exists $r_o > 0$ such that

$$c/2 \le \det D\mathbf{u}_r(\mathbf{x}) \le 2/c \tag{92}$$

for almost every $\mathbf{x} \in \Omega$ and for every $|r| < r_o$. Thanks to Remark 4.1, since $\mathbf{u}_r \in C^1(\Omega)^d \cap C(\bar{\Omega})^d$, $\mathbf{u}_r$ and $\bar{\mathbf{u}}$ agree on $\partial\Omega$, (92) implies that $\mathbf{u}_r$ is one-to-one from $\bar{\Omega}$ onto $\bar{\mathbf{u}}(\bar{\Omega})$ and $\mathbf{u}_r \in \mathcal{U}_\Lambda$. Using that $\bar{\mathbf{u}}$ maximizes I over $\mathcal{U}_\Lambda$ we have that

$$0 = -\lim_{r\to 0}(I[\mathbf{u}_r] - I[\bar{\mathbf{u}}])/r = \lim_{r\to 0}\int_K (W(D\mathbf{u}_r) - W(D\bar{\mathbf{u}}))/r d\mathbf{x} - \int_K \mathbf{F}\cdot\mathbf{v} d\mathbf{x}. \tag{93}$$

Since $(D\mathbf{u}_r)$ converges uniformly to $D\bar{\mathbf{u}}$ on Ω, $\{AdjD\mathbf{u}_r\}_r$ and $AdjD\bar{\mathbf{u}}$ are uniformly bounded by a constant $c_1 > 0$. Now note that DW is bounded on $\{M \in \mathbf{R}^{d\times d} : c/2 \le \det M \le 2/c, |AdjM| < c_1\}$, and so (92) and (93) yield

$$0 = \int_K DW(D\bar{\mathbf{u}}) \cdot D\mathbf{v} d\mathbf{x} = \int_K \mathbf{F}\cdot\mathbf{v} d\mathbf{x}. \tag{94}$$

Since $\mathbf{v}$ is arbitrary in (94) we read off

$$-\operatorname{div}(DW(D\bar{\mathbf{u}})) = \mathbf{F} \qquad \text{in } \Omega$$

in the weak sense.

This concludes the proof of Theorem 4.1. □

Remark 4.1. If $\mathbf{u}_o \in C^1(\Omega)^d \cap C(\bar{\Omega})^d$ is one-to-one on Ω, $\det D\mathbf{u}_o$ is positive, and $\mathbf{u}_o(\Omega) := \Lambda$, then by the invariance of domain theorem the set Λ is open (see [16]). If $\mathbf{u} \in C^1(\Omega)^d \cap C(\bar{\Omega})^d$ agrees with $\mathbf{u}_o$ on $\partial\Omega$, then

$$deg(\mathbf{u}, \Omega, \mathbf{y}) = deg(\mathbf{u}_o, \Omega, \mathbf{y}) = \begin{cases} 1 & \text{if} \quad \mathbf{y} \in \Lambda, \\ 0 & \text{if} \quad \mathbf{y} \notin \bar{\Lambda}, \end{cases} \tag{95}$$

where $deg(\mathbf{u}, \Omega, \mathbf{y})$ stands for the topological degree of $\mathbf{u}$ at $\mathbf{y}$ on Ω. If in addition $\det D\mathbf{u} > 0$ in Ω, then (95) implies $\mathbf{u}$ is one-to-one and $\mathbf{u}(\Omega) = \Lambda$. Hence $\mathbf{u} \in \mathcal{U}_\Lambda$. In particular, $\mathcal{U}_o$ is a subset of $\mathcal{U}_\Lambda$ (see, for instance, [16] for properties of the topological degree theory).

COROLLARY 4.2. *Assume that* $\mathbf{u}_o \in C^1(\Omega)^d \cap C(\bar{\Omega})^d$ *is one-to-one on* $\bar{\Omega}$, $\det D\mathbf{u}_o$ *is positive and belongs to* $C^1(\Omega)$, $\det D\mathbf{u}_o + 1/\det D\mathbf{u}_o \in L^\infty(\Omega)$, *and* $\mathbf{u}_o(\Omega) = \Lambda$. *Under the assumptions of Theorem* 4.1 *the infima in* (7) *and* (8) *coincide.*

Proof. Thanks to Remark 4.1 we have that $\inf_{\mathcal{U}_\Lambda} E \le \inf_{\mathcal{U}_o} E$. To conclude the proof of the corollary it suffices to show the reverse inequality. Let $\bar{\mathbf{u}}$ be the minimizer of E over $\mathcal{U}_\Lambda$. By Proposition C.1 there exists a sequence $(\mathbf{u}_n) \subset \mathcal{U}_o$ such that $||\mathbf{u}_n - \bar{\mathbf{u}}||_1 ||\mathbf{F}||_\infty \le 1/n$ and

$$\det D\mathbf{u}_n = \det D\bar{\mathbf{u}} \qquad \text{almost everywhere in } \Omega$$

for each $n = 1, 2, \ldots$. We have that

$$E[\mathbf{u}_n] = E[\bar{\mathbf{u}}] + \int_\Omega \mathbf{F} \cdot (\bar{\mathbf{u}} - \mathbf{u}_n) d\mathbf{x} \le \inf_{\mathcal{U}_\Lambda} E + 1/n.$$

This concludes the proof of Corollary 4.2. □

Appendix A. Properties of the map $\phi \to \phi^\sharp$. Throughout this section Λ is an open subset of $\mathbf{R}^d$ contained in the closed ball B_R of center 0 and radius $R > 0$, $h \in C^2(0, +\infty)$ is strictly convex and satisfies the growth conditions (4). Recall that

$$H(t) := h(t) - th'(t) \qquad (t \in (0, +\infty)).$$

Suppose that $\tilde{\phi} : conv(\Lambda) \to \mathbf{R}$, $\tilde{\psi} : \mathbf{R}^d \to \mathbf{R} \cup \{+\infty\}$ are lower semicontinuous, and define the convex functions

$$\psi(\mathbf{z}) = \tilde{\phi}^\sharp(\mathbf{z}) := \sup_{\mathbf{y} \in conv(\Lambda)} \{\mathbf{y} \cdot \mathbf{z} + h^*(-\tilde{\phi}(\mathbf{y}))\} \qquad (\mathbf{z} \in \mathbf{R}^d), \tag{96}$$

and

$$\phi(\mathbf{y}) = \tilde{\psi}_\sharp(\mathbf{y}) := \sup_{\alpha > 0} \left\{ \frac{(\tilde{\psi})^*(\mathbf{y}) - h(\alpha)}{\alpha} \right\} \qquad (\mathbf{y} \in \mathbf{R}^d). \tag{97}$$

LEMMA A.1. *Let* $\mathbf{y}_o, \mathbf{z}_o \in \mathbf{R}^d$. *The following statements hold:*

(i) *The supremum in* $\phi(\mathbf{y}_o)$ *is attained for* $\beta(\mathbf{y}_o) \in (0, +\infty)$ *provided that* $(\tilde{\psi})^*(\mathbf{y}_o)$ *is finite. If* $S(\mathbf{y}_o) \in \partial(\tilde{\psi})^*(\mathbf{y}_o)$, *then we have that* $S(\mathbf{y}_o) \in \beta(\mathbf{y}_o)\partial\phi(\mathbf{y}_o)$, *and* $H(\beta(\mathbf{y}_o))$

$= (\tilde{\psi})^*(\mathbf{y}_o)$. *Consequently, the pair* $(\beta(\mathbf{y}_o), S(\mathbf{y}_o))$ *or in other words the pair* $(\beta(\mathbf{y}_o), D\tilde{\psi}^*(\mathbf{y}_o))$ *is uniquely determined if* $(\tilde{\psi})^*$ *is differentiable at* $\mathbf{y}_o$; β, S *are Borel functions.*

(ii) *If* $\psi \not\equiv +\infty$, *then* $Lip(\psi) \leq R$.

(iii) *If* $(\tilde{\psi}, \tilde{\phi}) \in \mathcal{A}$, *then* $\phi \leq \tilde{\phi}$ *on* $conv(\Lambda)$ *and* $\psi \leq \tilde{\psi}$ *on* $\mathbf{R}^d$.

(iv) *We have that* $((\tilde{\phi}^\sharp)_\sharp)^\sharp = \tilde{\phi}^\sharp$ *on* $\mathbf{R}^d$ *and* $((\tilde{\psi}_\sharp)^\sharp)_\sharp = \tilde{\psi}_\sharp$ *on* $conv(\Lambda)$.

Proof. Step 1. We first prove (i). Note that in light of (4), $\phi(\mathbf{y}_o)$ is finite if and only if $(\tilde{\psi})^*(\mathbf{y})$ is finite, in which case existence of a maximizer $\beta(\mathbf{y}_o)$ in $\phi(\mathbf{y}_o)$ is a straightforward to obtain. Next, observe that if $S(\mathbf{y}_o) \in \partial(\tilde{\psi})^*(\mathbf{y}_o)$, then the auxiliary function $K : (\alpha, \mathbf{y}, \mathbf{z}) \to \alpha\phi(\mathbf{y}) + \tilde{\psi}(\mathbf{z}) + h(\alpha) - \mathbf{y} \cdot \mathbf{z}$ attains its minimum at $(\beta(\mathbf{y}_o), \mathbf{y}_o, S(\mathbf{y}_o))$. Exploiting the fact that both functions K and $\frac{\partial K}{\partial \alpha}$ vanish at $(\beta(\mathbf{y}_o), \mathbf{y}_o, S(\mathbf{y}_o))$ we deduce that

$$-\phi(\mathbf{y}_o) = h'(\beta(\mathbf{y}_o)) \text{ and } H(\beta(\mathbf{y}_o)) = \mathbf{y}_o \cdot S(\mathbf{y}_o) - \tilde{\psi}(S(\mathbf{y}_o)).$$

Step 2. Since $K(\beta(\mathbf{y}_o), \mathbf{y}_o, \mathbf{z})$ and $K(\beta(\mathbf{y}_o), \mathbf{y}, S(\mathbf{y}_o))$ are greater than or equal to $K(\beta(\mathbf{y}_o), \mathbf{y}_o, S(\mathbf{y}_o))$, we readily deduce that $S(\mathbf{y}_o) \in \beta(\mathbf{y}_o)\partial\phi(\mathbf{y}_o)$. Using the fact that $\tilde{\psi}(S(\mathbf{y}_o)) + (\tilde{\psi})^*(\mathbf{y}_o) = \mathbf{y}_o \cdot S(\mathbf{y}_o)$, the equation $H(\beta(\mathbf{y}_o)) = \mathbf{y}_o \cdot S(\mathbf{y}_o) - \tilde{\psi}(S(\mathbf{y}_o))$ reads off $H(\beta(\mathbf{y}_o)) = (\tilde{\psi})^*(\mathbf{y}_o)$. This concludes the proof of (i). Since $\Lambda \subset B_R$ we conclude (ii).

Step 3. The proof of (iii) is straightforward.

Step 4. We now prove (iv). We have that $(\tilde{\phi}^\sharp, (\tilde{\phi}^\sharp)_\sharp) \in \mathcal{A}$ and because $(\tilde{\phi}^\sharp, \tilde{\phi}) \in \mathcal{A}$, (iii) implies that $(\tilde{\phi}^\sharp)_\sharp \leq \tilde{\phi}$ on $conv(\Lambda)$. Using the fact that the operator $\varphi \to \varphi^\sharp$ is nonincreasing we deduce that $((\tilde{\phi}^\sharp)_\sharp)^\sharp \geq \tilde{\phi}^\sharp$ on $\mathbf{R}^d$. But (iii) and $(\tilde{\phi}^\sharp, (\tilde{\phi}^\sharp)_\sharp) \in \mathcal{A}$ also imply that $((\tilde{\phi}^\sharp)_\sharp)^\sharp \leq \tilde{\phi}^\sharp$ on $\mathbf{R}^d$. Consequently, $((\tilde{\phi}^\sharp)_\sharp)^\sharp = \tilde{\phi}^\sharp$ on $\mathbf{R}^d$. Likewise, $((\tilde{\psi}_\sharp)^\sharp)_\sharp = \tilde{\psi}_\sharp$ on $conv(\Lambda)$.

This concludes the proof of Lemma A.1. □

LEMMA A.2. *Suppose that* $\tilde{\psi} \equiv +\infty$ *on the complement of* B_R *and that* $|\tilde{\psi}|_{L^\infty(B_R)} < +\infty$. *Let* β *be defined as in Lemma* A.1. *Then there exists a constant* c *depending only on* h, R, *and* $|\tilde{\psi}|_{L^\infty(B_R)}$ *such that* $c \leq \beta(\mathbf{y}) \leq 1/c$ *for all* $\mathbf{y}$.

Proof. Set $t_o := R^2 + |\tilde{\psi}|_{L^\infty(B_R)}$. Since $\tilde{\psi} \equiv +\infty$ on the complement of B_R we obtain that $|(\tilde{\psi})^*|_{L^\infty(B_R)} \leq t_o$. Using (12) and Lemma A.1 (i) we conclude the lemma with $c := \max\{H^{-1}(t_o), 1/H^{-1}(-t_o)\}$. □

Appendix B. Compacity of a special class of measures. Throughout this section we assume that $\Lambda \subset \mathbf{R}^d$ is an open bounded set whose closure is contained in the closed ball B_{R_o} of center 0 and radius R_o. If μ is a finite positive measure on $\mathbf{R}^d$, we recall that the moments $M_o(\mu)$ and $M_o(\mu)$ are defined in (20), $C := (0, \infty) \times \mathbf{R}^d \times \mathbf{R}^d$, and $\Gamma[\mu]$ is the set of all Borel measures on C such that

$$\int_C f(\mathbf{z}) d\gamma(\alpha, \mathbf{y}, \mathbf{z}) = \int_{\mathbf{R}^d} f(\mathbf{z}) d\mu(\mathbf{z})$$

and

$$\int_C \alpha f(\mathbf{y}) d\gamma(\alpha, \mathbf{y}, \mathbf{z}) = \int_\Lambda f(\mathbf{y}) d\mathbf{y}$$

for all $f \in C_o(\mathbf{R}^d)$.

PROPOSITION B.1. *Suppose that* μ *satisfies* (20), *that* (μ_n) *is a sequence of Borel measures converging weak* $*$ *to* μ, $M_o[\mu_n] = M_o[\mu]$ $(n = 1, 2, \ldots)$, *and that*

h satisfies (4). *If $\gamma_n \in \Gamma(\mu_n)$ and the sequence of real numbers $(\int_C |h(\alpha)|d\gamma_n)$ is bounded independently of n, then there exists a sequence $(n_j) \subset \mathbf{N}$ and a Borel measure $\gamma \in \Gamma(\mu)$ such that (γ_{n_j}) converges weak $*$ to γ.*

Proof. Because $\gamma_n \in \Gamma(\mu_n)$ we have that $\gamma_n[C] = M_o[\mu]$, and so there exists a sequence $(n_j) \subset \mathbf{N}$ and a Borel measure γ on C such that (γ_{n_j}) converges weak $*$ to γ. We next introduce the functions

$$k(\alpha, \mathbf{y}) := l_R(\alpha + |\mathbf{y}|) \qquad (\alpha > 0, \mathbf{y} \in \mathbf{R}^d),$$

where, for $R > 1$, $l_R : \mathbf{R} \to [0, 1]$ is of class C^∞ and satisfies

$$l_R(t) = \begin{cases} 1 & \text{if} \quad |t| \le R - 1, \\ 0 & \text{if} \quad |t| \ge R. \end{cases} \tag{98}$$

If $f \in C_o(\mathbf{R}^d)$, then

$$\begin{aligned} \left| \int_C f(\mathbf{z})(1 - k(\alpha, \mathbf{y}))d\gamma_{n_j} \right| &= \left| \int_{\alpha > (R-1)/2} f(\mathbf{z})(1 - k(\alpha, \mathbf{y}))d\gamma_{n_j} \right| \\ &\le 2(|f|_\infty |\Lambda|)/(R - 1). \end{aligned} \tag{99}$$

Using (99) and the fact that $\gamma_{n_j} \in \Gamma(\mu_{n_j})$ we have that

$$\left| \int_{\mathbf{R}^d} f(\mathbf{z})d\mu_{n_j}(\mathbf{z}) - \int_C f(\mathbf{z})k(\alpha, \mathbf{y})d\gamma_{n_j} \right| \le 2(|f|_\infty |\Lambda|)/(R - 1). \tag{100}$$

Letting first j go to $+\infty$ and then R go to $+\infty$ in (100) we deduce that

$$\int_{\mathbf{R}^d} f(\mathbf{z})d\mu(\mathbf{z}) = \int_C f(\mathbf{z})d\gamma. \tag{101}$$

Define the function

$$\beta(R) := M \sup_t \{t/|h(t)| \mid t \ge (R - 1)/2\} \qquad (R > 1),$$

where $M > 0$ is a constant independent of n such that $\int_C |h(\alpha)|d\gamma_n \le M$ for all $n \in \mathbf{N}$. Since $\gamma_{n_j} \in \Gamma(\mu_{n_j})$, if A_R is the subset of all $(\alpha, \mathbf{y}, \mathbf{z}) \in C$ such that $|\mathbf{z}| > (R - 1)/2$ and $|\alpha| \le (R - 1)/2$, then we have that

$$\left| \int_\Lambda f(\mathbf{y})d\mathbf{y} - \int_C \alpha f(\mathbf{y})k(\alpha, \mathbf{z})d\gamma_{n_j} \right| \le \left| \int_C \alpha f(\mathbf{y})(1 - k(\alpha, \mathbf{z}))d\gamma_{n_j} \right| \tag{102}$$

and

$$\begin{aligned} \left| \int_C \alpha f(\mathbf{y})(1 - k(\alpha, \mathbf{z}))d\gamma_{n_j} \right| &\le 2 \int_{\alpha > (R-1)/2} \alpha |f(\mathbf{y})|(1 - k(\alpha, \mathbf{z}))d\gamma_{n_j} \\ &\quad + \int_{A_R} \alpha |f(\mathbf{y})|(1 - k(\alpha, \mathbf{z}))d\gamma_{n_j} \\ &\le 2|f|_\infty \Big(\beta(R) + R(\mu[B^c_{\frac{R-1}{2}}] + 1/n_j) \Big). \end{aligned}$$

Hence

$$\left| \int_C \alpha f(\mathbf{y})(1 - k(\alpha, \mathbf{z}))d\gamma_{n_j} \right| \le 2|f|_\infty \Big(\beta(R) + \int_{B^c_{\frac{R-1}{2}}} (2|\mathbf{z}| + 1)d\mu + R/n_j \Big). \tag{103}$$

In light of (4) $\beta(R)$ tends to 0 as R tends to $+\infty$. Using (102) and letting first j go to $+\infty$ and then R tend to $+\infty$ in (103), since $M_o[\mu], M_1[\mu] < +\infty$, we deduce that

$$\int_\Lambda f(\mathbf{y})d\mathbf{y} = \int_C \alpha f(\mathbf{y})d\gamma. \tag{104}$$

Since $f \in C_o(\mathbf{R}^d)$ is arbitrary (101) and (104) yield $\gamma \in \Gamma(\mu)$, which concludes the proof of Proposition B.1. □

Appendix C. Density of the set of maps with prescribed boundary values.

PROPOSITION C.1. *Suppose that $d \geq 2$, $\Omega, \Lambda \subset \mathbf{R}^d$ are two open, bounded sets, that $\partial\Omega$ is Lipschitz, and that Λ is convex. Let $\mathbf{u}, \mathbf{u}_o \in C^1(\Omega)^d \cap C(\bar{\Omega})^d$ be such that $\det D\mathbf{u}, \det D\mathbf{u}_o$ are positive, of class $C^1(\Omega)$, with $\det D\mathbf{u} + \frac{1}{\det D\mathbf{u}}$ and $\det D\mathbf{u}_o + \frac{1}{\det D\mathbf{u}_o}$ in $L^\infty(\Omega)$. Suppose furthermore that $\mathbf{u}_o$ is one-to-one on $\bar{\Omega}$, that $\mathbf{u}$ is one-to-one on Ω, and that $\mathbf{u}(\Omega) = \mathbf{u}_o(\Omega) = \Lambda$. Then there exists a sequence $(\mathbf{u}_n) \subset C^1(\Omega)^d \cap C(\bar{\Omega})^d$ of one-to-one maps from $\bar{\Omega}$ onto $\bar{\Lambda}$ converging almost everywhere in Ω to $\mathbf{u}$ and such that for each integer n*

$$\begin{cases} \det D\mathbf{u}_n = \det D\mathbf{u} & \text{almost everywhere in } \Omega, \\ \mathbf{u}_n = \mathbf{u}_o & \text{on } \partial\Omega. \end{cases} \tag{105}$$

Proof. Step 1. Using Theorem 7 in [9] we find $\mathbf{b} \in Diff^1(\Omega) \cap Diff^0(\bar{\Omega})$ such that

$$\begin{cases} \det D\mathbf{u}_0(\mathbf{b}(\mathbf{x}))\det D\mathbf{b}(\mathbf{x}) = \det D\mathbf{u}(\mathbf{x}) & \text{in } \Omega, \\ \mathbf{b}(\mathbf{x}) = \mathbf{x} & \text{on } \partial\Omega. \end{cases} \tag{106}$$

Define the maps

$$\mathbf{v} := \mathbf{u}_o \circ \mathbf{b}, \qquad \mathbf{s} := \mathbf{u} \circ \mathbf{v}^{-1}.$$

Clearly

$$\begin{cases} \det D\mathbf{v} = \det D\mathbf{u} & \text{in } \Omega, \\ \mathbf{v}(\mathbf{x}) = \mathbf{u}_o(\mathbf{x}) & \text{on } \partial\Omega. \end{cases} \tag{107}$$

We have that

$$\mathbf{s}(\Lambda) = \mathbf{u}(\mathbf{b}^{-1}[\mathbf{u}_o^{-1}(\Lambda)]) = \mathbf{u}(\Omega) = \Lambda$$

and $\mathbf{s}$ is measure-preserving in the sense that

$$\int_\Lambda G(\mathbf{s}(\mathbf{y}))d\mathbf{y} = \int_\Lambda G(\mathbf{x})d\mathbf{x}$$

for all $G \in C_o(\mathbf{R}^d)$.

Step 2. Since Λ is convex and bounded, there exists a map $T \in Diff^1(\Lambda, (0,1)^d) \cap Diff^0(\bar{\Lambda}, [0,1]^d)$. One can choose T, for instance, to be the optimal map that rearranges $\frac{\chi_\Lambda}{|\Lambda|}d\mathbf{x}$ onto $\chi_{[0,1]^d}d\mathbf{x}$ in the Monge problem, where optimality is measured against the cost function $c(\mathbf{x}-\mathbf{y}) = |\mathbf{x}-\mathbf{y}|^2$. Using T we deduce that the following known result for $[0,1]^d$ (see, for instance, [2] and [30]) holds for any convex, bounded

set Λ: there exists a sequence $(\mathbf{s}_n) \subset C^1(\Lambda)^d \cap C(\bar{\Lambda})^d$ of maps from $\bar{\Lambda}$ onto $\bar{\Lambda}$ that are one-to-one on Λ, that converge pointwise almost everywhere in Λ to $\mathbf{s}$ such that

$$\begin{cases} \det D\mathbf{s}_n &= 1 \quad \text{in} \quad \Lambda, \\ \mathbf{s}_n(\mathbf{y}) &= \mathbf{y} \quad \text{on} \quad \partial\Lambda \end{cases} \tag{108}$$

for $n = 1, 2, \ldots$. Define

$$\mathbf{u}_n(\mathbf{x}) := \mathbf{s}_n(\mathbf{v}(\mathbf{x})) \qquad (\mathbf{x} \in \bar{\Omega}).$$

By (107) and (108) we deduce that $(\mathbf{u}_n)$ satisfies the conclusions of Proposition C.1. □

Appendix D. Background on the Monge problem. In this section we present a brief description of the Monge problem, a theory which has attracted a lot of attention. Throughout this section we keep our focus only on the case that is relevant to the study of solid crystals, the case studied by [3], [19], etc. Let $\mu = f d\mathbf{x}$, $\nu = g d\mathbf{x}$ be finite measures on $\mathbf{R}^d$ with equal total mass. Let $O_1, O_2 \subset \mathbf{R}^d$ be two open sets such that $\bar{O}_1$ is the support of μ and $\bar{O}_2$ is the support of ν. The Monge mass transport problem consists of finding an optimal way of rearranging μ onto ν against a cost function which we choose here to be $c(\mathbf{x} - \mathbf{y}) = |\mathbf{x} - \mathbf{y}|^2$. The corresponding variational problem is to minimize the total work

$$K[T] := \int_{\mathbf{R}^d} |\mathbf{x} - T\mathbf{x}|^2 d\mu(\mathbf{x})$$

over the set $\mathcal{T}$ of all Borel maps $T : \mathbf{R}^d \to \mathbf{R}^d$ that push μ forward to ν. Define

$$K'[S] := \int_{\mathbf{R}^d} |\mathbf{y} - S\mathbf{y}|^2 d\nu(\mathbf{y})$$

and let $\mathcal{S}$ be the set of all Borel maps $S : \mathbf{R}^d \to \mathbf{R}^d$ that push ν forward to μ. The following results are known in a setting more general than the one herein.

PROPOSITION D.1 (general theorem).

(i) *Existence and uniqueness of optimal maps: there exists a unique T_o minimizing K over $\mathcal{T}$. Likewise, there exists a unique S_o minimizing K' over $\mathcal{S}$. We have that $S_o(T_o(\mathbf{x})) = \mathbf{x}$ for μ-almost every $\mathbf{x} \in \mathbf{R}^d$, $T_o(S_o(\mathbf{y})) = \mathbf{y}$ for ν-almost every $\mathbf{y} \in \mathbf{R}^d$.*

(ii) *Characterization of optimal maps: a map T_o is a minimizer of K over $\mathcal{T}$ if and only if $T_o \in \mathcal{T}$ and T_o is the gradient of a convex function $\psi_o : \mathbf{R}^d \to \mathbf{R} \cup \{+\infty\}$. Similarly, a map S_o is a minimizer of K' over $\mathcal{S}$ if and only if $S_o \in \mathcal{S}$ and S_o is the gradient of a convex function $\phi_o : \mathbf{R}^d \to \mathbf{R} \cup \{+\infty\}$.*

(iii) *The sets $T_o(O_1)$ and O_2 coincide up to a set of zero measure.*

Proof. We refer the reader to [19]. □

PROPOSITION D.2 (smoothness of optimal maps). *Assume that O_1, O_2 are bounded, $|\partial O_1| = |\partial O_2| = 0$, $f + 1/f \in L^\infty(O_1)$, $g + 1/g \in L^\infty(O_2)$, O_2 is convex, and ψ_o, ϕ_o are the convex functions obtained in Proposition* D.1. *Then we have the following:*

(i) *$\psi_o \in C^{1,s}(O_1)$ for some $0 < s < 1$, and ψ_o is strictly convex in O_1.*

(ii) *If in addition O_1 is convex, then $\psi_o \in C^{1,s}(\bar{O}_1)^d$ for some $0 < s < 1$.*

(iii) *If O_1 is convex and in addition $f \in C^{0,\bar{s}}(O_1)$, $g \in C^{0,\bar{s}}(O_2)$, then $D\psi_o \in C^{1,s}(O_1)^d \cap C^{0,\bar{s}}(\bar{O}_1)^d$, $D\phi_o \in C^{1,s}(O_2)^d \cap C^{0,\bar{s}}(\bar{O}_2)^d$ for all $0 < s < \bar{s}$. We have that $D\psi_o \in Diff^0(\bar{O}_1, \bar{O}_2)$.*

Proof. Smoothness properties of ψ_o and ϕ_o as stated in (i), (ii), and (iii) are established in [4], [5], and [6]. If $D\psi_o \in C^{0,\bar{s}}(\bar{O}_1)^d$ and $D\phi_o \in C^{0,\bar{s}}(\bar{O}_2)^d$, then by Proposition D.1 we have that $D\psi_o \in Diff^0(\bar{O}_1, \bar{O}_2)$. □

Acknowledgments. The authors would like to thank B. Dacorogna, D. Kinderlehrer, and A. Swiech for comments on the paper.

REFERENCES

[1] G. Alberti, *On the structure of singular sets of convex functions*, Calc. Var. Partial Differential Equations, 2 (1994), pp. 17–27.
[2] Y. Brenier, *personal notes.*
[3] Y. Brenier, *Décomposition polaire et réarrangement monotone des champs de vecteurs*, C.R. Acad. Sci. Paris Sér. I Math., 305 (1987), pp. 805–808.
[4] L.A. Caffarelli, *The regularity of mappings with a convex potential*, J. Amer. Math. Soc., 5 (1992), pp. 99–104,.
[5] L. Caffarelli, *Boundary regularity of maps with convex potentials*, Comm. Pure Appl. Math., 45 (1992), pp. 1141–1151.
[6] L.A. Caffarelli, *Boundary regularity of maps with convex potentials.* II, Ann. of Math. (2), 144 (1996), pp. 453–496.
[7] M. Chipot and D. Kinderlehrer, *Equilibrium configurations of crystals*, Arch. Rational Mech. Anal. , 103 (1988), pp. 237–277.
[8] B. Dacorogna, *Direct Methods in the Calculus of Variations*, Springer-Verlag, Berlin, 1989.
[9] B. Dacorogna and J. Moser, *On a partial differential equation involving the Jacobian determinant*, Ann. Inst. H. Poincare Anal. Non Linéaire, 7 (1990), pp. 1–26.
[10] I. Ekeland and R. Teman, *Convex Analysis and Variational Problems*, Stud. Math. Appl. 1, North-Holland, Amsterdam, 1976.
[11] J.L. Ericksen, *Special topics in electrostatics*, Advance in Applied Math. Mech., 17 (1977), pp. 188–224.
[12] J.L. Ericksen, *Twinning of crystals*, in Metastability and Incompletely Posed Problems, S. Antman, J.L. Ericksen, D. Kinderlehrer, and I. Müller, eds., Springer-Verlag, Berlin, 1987, pp. 77–94.
[13] L.C. Evans and R.F. Gariepy, *Measure Theory and Fine Properties of Functions*, Stud. Adv. Math., CRC Press, Boca Raton, FL, 1992.
[14] I. Fonseca, *Variational methods for elastic crystals*, Arch. Rational Mech. Anal., 97 (1987), pp. 189–220.
[15] I. Fonseca, *The lower quasiconvex envelope of the stored energy function for an elastic crystals*, J. Math. Pures Appl. (9), 67 (1988), pp. 175–195.
[16] I. Fonseca and W. Gangbo, *Degree Theory in Analyis and Applications*, Clarendon Press, Oxford, UK, 1995.
[17] I. Fonseca and L. Tartar, *The displacement problem for elastic crystals*, Proc. Royal Soc. Edinburgh Sect. A, 113 (1989), pp. 159–180.
[18] W. Gangbo, *An elementary proof of the polar factorization of vector-valued functions*, Arch. Rational Mech. Anal., 128 (1994), pp. 381–399.
[19] W. Gangbo and R.J. McCann, *The geometry of optimal transportation*, Acta Math., 177 (1996), pp. 113–161.
[20] R.D. James, *Mechanics of Coherent Phase Transformations in Solids*, MRL Report, Brown University, Division of Engineering, 1982, pp. 185–211.
[21] R.D. James, *The arrangement of coherent phases in a loaded body*, in Phase Transformations and Material Instabilities in Solids, M. Gurtin, ed., Academic Press, New York, 1984, pp. 79–98.
[22] R.D. James, *Stress free joints and polycrystals*, Arch. Rational Mech. Anal., 86 (1984), pp. 13–37.
[23] R.D. James, *Displacive phase transformations in solids*, J. Mech. Phys. Solids, 34 (1986), pp. 359–394.
[24] D. Kinderlehrer, *Twinning of crystals.* II, in Metastability and Incompletely Posed Problems, S. Antman, J.L. Ericksen, D. Kinderlehrer, and I. Müller, eds., Springer-Verlag, Berlin, 1987, pp. 185–211.
[25] G. Parry, *One phase transitions involving internal strain*, Internat. J. Solids Structures, 17 (1981), pp. 361–378.

[26] M. Pitteri, *Reconciliation of local and global symmetries of crystals*, J. Elasticity, 14 (1984), pp. 175–190.
[27] M. Pitteri, *On the kinematics of mechanical twinning in crystals*, Arch. Rational Mech. Anal., 88 (1985), pp. 25–57.
[28] M. Pitteri, *On type-2 twins in crystals*, Internat. J. Plasticity, 2 (1986), pp. 99–106.
[29] M. Pitteri, *A contribution to the description of natural states for elastic crystalline solids*, in Metastability and Incompletely Posed Problems, IMA Vol. Math. Appl. 3, S. Antman, J.L. Ericksen, D. Kinderlehrer, and I. Müller, eds., Springer-Verlag, Berlin, 1987, pp. 295–310.
[30] M. Roesch, Thèse de Doctorat, Université Pierre et Marie Curie, Paris VI, 1995.

SIAM J. MATH. ANAL.
Vol. 32, No. 3, pp. 493–503

VORTEX STATE OF d-WAVE SUPERCONDUCTORS IN THE GINZBURG–LANDAU ENERGY *

FANGHUA LIN† AND TAI-CHIA LIN‡

Abstract. We find a minimizer of a reduced form of the Ginzburg–Landau free energy for d-wave superconductors having distinct degree-one vortices. For a single vortex in the vortex core, we analytically recover the vortex structure with fourfold symmetry.

Key words. Ginzburg–Landau energy, d-wave, superconductor, vortex

AMS subject classifications. 35J35, 82D55

PII. S0036141099353527

1. Introduction.

In the 1910s, low-temperature superconductivity was observed on metals and alloys (cf. [9]). Recently, high-temperature superconductivity has been found on some copper-oxide superconductors (cf. [12]). The vortex state of high-temperature superconductors is different from the vortex state of low-temperature superconductors. When the applied magnetic field is close to the lower critical field H_{c_1}, the single vortex is expected to be symmetric in low-temperature superconductors but it may be asymmetric (fourfold symmetric) in high-temperature superconductors (cf. [8], [31]). Moreover, as the applied magnetic field is close to the upper critical field H_{c_2}, Abrikosov type vortex lattices are expected to be triangular in low-temperature superconductors but they may be rectangular in high-temperature superconductors (cf. [1], [8], [27], [30], [31], etc).

To distinguish low-temperature and high-temperature superconductivity, an s-wave and a d-wave order parameter were introduced (cf. [13], [21]). Soininen et al. [3] and Soininen, Kallin, and Berlinsky [28] introduced the Ginzburg–Landau free energy with an s-wave and a d-wave order parameter. Ren, Xu, and Ting (cf. [24], [25]) present a microscopic derivation of the Ginzburg–Landau equations from the Gor'kov equations by using the finite temperature Green's-function approximation method. From [31], we learned the two fields Ginzburg–Landau free energy is given by

$$\begin{aligned} G(\Psi_s,\Psi_d,A) = \int_{\mathbb{R}^2} & \kappa^2|\ \mathrm{curl}\ A-H|^2+\alpha_s(T)|\Psi_s|^2 \\ & +\tfrac{1}{2}(1-|\Psi_d|^2)^2+\tfrac{4}{3}|\Psi_s|^4+\tfrac{8}{3}|\Psi_s|^2|\Psi_d|^2+\tfrac{2}{3}(\Psi_s^2\Psi_d^{*^2}+\Psi_d^2\Psi_s^{*^2}) \\ & +2|\textstyle\prod\Psi_s|^2+|\prod\Psi_d|^2+\{\prod_x\Psi_s\prod_x^*\Psi_d^*-\prod_y\Psi_s\prod_y^*\Psi_d^*+\mathrm{H.C.}\}\,, \end{aligned} \tag{1.1}$$

where Ψ_s is the s-wave order parameter, Ψ_d is the d-wave order parameter and A is the vector-valued magnetic potential, $\prod = i\nabla - A$, H is a constant applied magnetic field, κ is the Ginzburg–Landau parameter, and

$$\alpha_s(T) = C_s/(1-T/T_c)\,. \tag{1.2}$$

*Received by the editors March 17, 1999; accepted for publication (in revised form) March 29, 2000; published electronically September 15, 2000. This work was partially supported by the National Science Center in Taiwan.

http://www.siam.org/journals/sima/32-3/35352.html

†Courant Institute of Mathematical Sciences, New York, NY (linf@cims.nyu.edu).

‡Department of Mathematics, Chung-Cheng University, Chia-Yi, Taiwan (tclin@math.ccu.edu.tw).

Here C_s is a positive constant, T is the current temperature, and T_c is the d-wave transition temperature.

As the current temperature T is close to T_c, Franz et al. [8] observed that in a predominantly d-wave superconductor, the s-wave component is generically very small. They also provided approximation formulas for the order parameters Ψ_d and Ψ_s as follows:

$$|\Psi_s| \ll |\Psi_d|, \quad |\nabla\Psi_s| \ll |\nabla\Psi_d| \quad \text{as } T \to T_c\,. \tag{1.3}$$

Affleck, Franz, and Amin [1] obtained the leading order in $(1 - T/T_c)$ as

$$\Psi_s = \xi \left(\prod{}_x^2 - \prod{}_y^2\right) \Psi_d\,, \tag{1.4}$$

where ξ is a parameter satisfying that $\xi \to 0$ as $T \to T_c$. In [7], Du derived (1.4) by the formal asymptotic analysis.

We learned from [5] and [6] that it is reasonable to ignore the magnetic field in strongly type II superconductors when the applied magnetic field is close to H_{c1} and $T \to T_c$. Hence it is valuable to study the two fields Ginzburg–Landau model (1.1) without the magnetic field (i.e., $A, H \equiv 0$). Moreover, Rosenstein et al. [6] took (1.3) and (1.4) into (1.1) and modified the free energy (1.1) as follows:

$$G(\Psi_d) = \int_{\mathbb{R}^2} |\nabla\Psi_d|^2 + \frac{1}{2}(1 - |\Psi_d|^2)^2 + \beta\,|\mathbb{B}\Psi_d|^2\,dx\,dy\,, \tag{1.5}$$

where $\mathbb{B} = \partial_x^2 - \partial_y^2$ and β is a parameter satisfying that $\beta \to 0$ as $T \to T_c$. Here we have ignored the magnetic field (i.e., $A, H \equiv 0$) for strongly type II superconductors.

It is hard to find the minimizer of (1.5) by the standard direct method. Suppose that $\Psi_d \in H^2(\mathbb{R}^2\,;\mathbb{C})$ is a minimizer of (1.5) over $H^2(\mathbb{R}^2\,;\mathbb{C})$. Then it is easy to check that

$$\begin{aligned} G(\Psi_d + v) = \;& G(\Psi_d) + \int_{\mathbb{R}^2} |\nabla v|^2 - (1 - |\Psi_d|^2)|v|^2 + 2(\Psi_d \cdot v)^2 \\ & + \int_{\mathbb{R}^2} 2|v|^2(\Psi_d \cdot v) + \frac{1}{2}|v|^4 + \beta|\mathbb{B}v|^2\,, \end{aligned} \tag{1.6}$$

for any test function $v \in C_0^\infty(\mathbb{R}^2)$. Hereafter, $z_1 \cdot z_2 = \frac{1}{2}(\bar{z_1}z_2 + z_1\bar{z_2})\ \forall\ z_1, z_2 \in \mathbb{C}$. Let $v_n(z) = \delta_n v_0(z)\sin[\delta_n^{-2/3}(x+y)]$ for $z = x + iy \in \mathbb{C} \cong \mathbb{R}^2$, where v_0 is a test function with a nonempty compact support and $\{\delta_n\}$ is a sequence of positive numbers such that $\delta_n \to 0$ as $n \to \infty$. Here we use the fact that the complex plane $\mathbb{C}$ is isomorphic to $\mathbb{R}^2$. Now we replace v in (1.6) by v_n and we obtain that $G(\Psi_d + v_n) \to G(\Psi_d)$ but $\|\Psi_d + v_n\|_{H^2} \to \infty$ as $n \to \infty$. Hence $\Psi_d + v_n$'s form a minimizing sequence but $\Psi_d + v_n$'s have no converging subsequence even weakly converging subsequences in $H^2_{loc}(\mathbb{R}^2\,;\mathbb{C})$. Thus the free energy (1.5) has a defect on minimization.

From [30], we learned a Ginzburg–Landau energy functional (without the magnetic field) as follows:

$$E(\Psi_d) = \int_{\mathbb{R}^2} |\nabla\Psi_d|^2 + \frac{1}{2}(1 - |\Psi_d|^2)^2 + \eta\,(|\partial_x^2\,\Psi_d|^2 + |\partial_y^2\,\Psi_d|^2)\,dx\,dy\,, \tag{1.7}$$

where η is a constant depending on the current temperature T. The term $|\partial_x^2\Psi_d|^2 + |\partial_y^2\Psi_d|^2$ breaks the circular symmetry and accounts for the square symmetry. Furthermore, Park and Huse [22] introduced a more generalized Ginzburg–Landau free

energy (without the magnetic field) for d-wave superconductors as follows:

$$F(\Psi_d) = \int_{\mathbb{R}^2} |\nabla\Psi_d|^2 + \frac{1}{2}(1-|\Psi_d|^2)^2 + \gamma_1 |\Delta\Psi_d|^2 \, dx\, dy + \int_{\mathbb{R}^2} \beta_1 \left(|\mathbb{B}\Psi_d|^2 - 4|\partial_x\partial_y \Psi_d|^2\right) dx\, dy\,, \tag{1.8}$$

where $\Delta = \partial_x^2 + \partial_y^2$ and β_1, γ_1 are parameters tending to zero as $T \to T_c$.

Hereafter, we assume that $|\Psi_d| \to 1$ and all the derivatives of Ψ_d decay fast as $|(x,y)| \to \infty$. Such an assumption is consistent with the results in [8] and [31]. Using integration by parts, we may transform (1.8) into

$$\tilde{G}(\Psi_d) = \int_{\mathbb{R}^2} |\nabla\Psi_d|^2 + \frac{1}{2}(1-|\Psi_d|^2)^2 + \beta\,|\mathbb{B}\Psi_d|^2 + \gamma\,|\Delta\Psi_d|^2 \, dx\, dy\,, \tag{1.9}$$

where β, γ are parameters tending to zero as $T \to T_c$. In this paper, we assume that $\beta, \gamma > 0$ and $\beta, \gamma \to 0$ as $T \to T_c$. In particular, such an assumption includes the case that $0 < \gamma \ll \beta$, i.e., (1.9) is a small perturbation of (1.5).

In section 2, we approximate (1.9) by

$$G_\epsilon(\Psi_d) = \int_{\frac{1}{\epsilon}\Omega} |\nabla\Psi_d|^2 + \frac{1}{2}(1-|\Psi_d|^2)^2 + \beta|\mathbb{B}\Psi_d|^2 + \gamma\,|\Delta\Psi_d|^2 \, dx\, dy\,, \tag{1.10}$$

where $0 < \epsilon \ll 1$ is a small parameter, Ω is a bounded smooth domain in $\mathbb{R}^2$ having an interior point at the origin, and $\frac{1}{\epsilon}\Omega = \{(\frac{x}{\epsilon}, \frac{y}{\epsilon}) : (x,y) \in \Omega\}$. In the rest of this paper, we prove that the minimizer of (1.10) has distinct degree-one vortices in section 3. In section 4, we replace $\frac{1}{\epsilon}\Omega$ in (1.10) by B_{R_0}, where B_{R_0} is a disk with radius R_0 and center at the origin. Here $R_0 > 0$ is a large constant satisfying $1 \ll R_0 \le 1/\epsilon$. Then (1.10) becomes

$$\hat{G}(\Psi_d) = \int_{B_{R_0}} |\nabla\Psi_d|^2 + \frac{1}{2}(1-|\Psi_d|^2)^2 + \beta|\mathbb{B}\Psi_d|^2 + \gamma\,|\Delta\Psi_d|^2 \, dx\, dy\,, \tag{1.11}$$

where $\beta > 0$ is a small parameter as $T \to T_c$, $\gamma = C\,\beta$, and C is a positive constant independent of β. We then study the critical point of (1.11) and find out its single vortex structure with fourfold symmetry. The single vortex structure of d-wave superconductors having fourfold symmetry is well known in physics (cf. [5], [6], [8], [27], and [31]). Here we give a mathematical proof of such a vortex structure.

2. Preliminaries. To investigate vortices in d-wave superconductors, we assume that the order parameter Ψ_d satisfies $|\Psi_d| \to 1$ and all the derivatives of Ψ_d decay fast as $|(x,y)| \to \infty$. Such an assumption is consistent with the results in [8] and [31]. Hence we may approximate (1.9) by

$$G_\epsilon(\Psi_d) = \int_{\frac{1}{\epsilon}\Omega} |\nabla\Psi_d|^2 + \frac{1}{2}(1-|\Psi_d|^2)^2 + \beta|\mathbb{B}\Psi_d|^2 + \gamma\,|\Delta\Psi_d|^2 \, dx\, dy\,, \tag{2.1}$$

where $0 < \epsilon \ll 1$ is a small parameter, Ω is a bounded smooth domain in $\mathbb{R}^2$ having an interior point at the origin, and $\frac{1}{\epsilon}\Omega = \{(\frac{x}{\epsilon}, \frac{y}{\epsilon}) : (x,y) \in \Omega\}$. Rescaling the spatial variables x, y by ϵ, (2.1) becomes

$$\hat{G}_\epsilon(\Psi_d) = \int_{\Omega} |\nabla\Psi_d|^2 + \frac{1}{2\epsilon^2}(1-|\Psi_d|^2)^2 + \delta_\epsilon|\mathbb{B}\Psi_d|^2 + \gamma_\epsilon\,|\Delta\Psi_d|^2 \, dx\, dy\,, \tag{2.2}$$

where

$$\delta_\epsilon = \beta\,\epsilon^2 \quad \text{and} \quad \gamma_\epsilon = \gamma\,\epsilon^2\,. \tag{2.3}$$

Of course, (2.3) implies that $0 < \delta_\epsilon, \gamma_\epsilon = O(\epsilon^2)$ as $\epsilon \to 0+$. In sections 2 and 3, we study (2.2) with an assumption that $0 < \delta_\epsilon, \gamma_\epsilon = O(\epsilon^2)$ as $\epsilon \to 0+$.

This kind of approximation can also be found in s-wave superconductors. The conventional s-wave Ginzburg–Landau free energy (cf. [9]) without the magnetic field is

$$\int_{\mathbb{R}^2} \frac{1}{2}|\nabla u|^2 + \frac{1}{4}(1-|u|^2)^2\,,$$

where $u \in \mathbb{C}$ is the s-wave order parameter. Under the hypothesis that $|u| \to 1$ and all the derivatives of u decay fast at $|(x,y)| \to \infty$, we may approximate the s-wave Ginzburg–Landau free energy by

$$\int_{\frac{1}{\epsilon}\Omega} \frac{1}{2}|\nabla u|^2 + \frac{1}{4}(1-|u|^2)^2\,,$$

where $0 < \epsilon \ll 1$ is a small parameter and Ω is a bounded smooth domain in $\mathbb{R}^2$ having an interior point at the origin. Then we rescale the spatial variables by ϵ and obtain the energy functional as follows:

$$E_\epsilon(u) = \int_\Omega \frac{1}{2}|\nabla u|^2 + \frac{1}{4\epsilon^2}(1-|u|^2)^2\,, \tag{2.4}$$

where $u : \Omega \to \mathbb{C}$ is the s-wave order parameter. There are many investigations on the free energy (2.4). For the readers who are interested in these works, please refer to [2], [15], [17], [23], [29], etc.

In [2] and [29], we learn the minimizer of E_ϵ over $H^1_g(\Omega)$ having n degree-one vortices in Ω, where

$$H^1_g(\Omega) = \{u \in H^1(\Omega\,;\,\mathbb{C}) \,:\, u = g \quad \text{on } \partial\Omega\}\,,$$

and $g : \partial\Omega \to S^1$ is smooth with degree $n \geq 1$. Furthermore, the minimizer u_ϵ of (2.4) satisfies

(1) $E_\epsilon(u_\epsilon) = n\pi \log \frac{1}{\epsilon} + W_g(a_1,\dots,a_n) + o_\epsilon(1)$ as $\epsilon \to 0+$,
(2) u_ϵ converges to u_* (up to a subsequence) in $C^2_{loc}(\bar\Omega\backslash\{a_1,\dots,a_n\})$ as $\epsilon \to 0+$,
(3) $(a_1,\dots,a_n) \in \Omega^n$ is a global minimizer of the renormalized energy W_g defined in [2],

where $o_\epsilon(1)$ is a small quantity which tends to zero as $\epsilon \to 0+$,

$$u_*(z) = \prod_{j=1}^n \frac{z-a_j}{|z-a_j|}\, e^{i\,h(z)} \quad \forall z \in \Omega\,, \tag{2.5}$$

and h is a real-valued harmonic function. Since $\mathbb{R}^2$ is isomorphic to $\mathbb{C}$, we may consider $\Omega \subset \mathbb{R}^2 \cong \mathbb{C}$. Note that the domain Ω is assumed star-shaped in [2]. However, Struwe [29] generalized results of [2] for all bounded smooth domains.

For the minimizer of (2.2), we prove the following theorem.

THEOREM 2.1. *Suppose* $0 < \delta_\epsilon, \gamma_\epsilon = O(\epsilon^2)$ *as* $\epsilon \to 0+$. *Then there exists a minimizer* u_ϵ *of* (2.2) *over* $H^1_g(\Omega)$ *such that*

(i) $u_\epsilon \in H^2(\Omega)$ *has* n *degree-one vortices in* Ω,
(ii) $\hat{G}_\epsilon(u_\epsilon) = 2n\pi \log \frac{1}{\epsilon} + O(1)$ *as* $\epsilon \to 0+$,
(iii) u_ϵ *converges to* u_* *(up to a subsequence) strongly in* $L^2(\Omega)$ *and weakly in* $H^1_{loc}(\Omega \backslash \{a_1, \dots, a_n\})$,
(iv) $(a_1, \dots, a_n) \in \Omega^n$ *is a global minimizer of the renormalized energy* W_g *in* [2].

Remark 1. We may consider the energy functional (2.2) with $0 < \delta_\epsilon, \gamma_\epsilon = O(\epsilon^2)$ as a small perturbation of (2.4). However, the perturbation terms are of higher order derivatives. Hence the Euler–Lagrange equation of (2.2) is a singular perturbation problem and the perturbation terms are of the 4th order derivatives. Until now, there is no general theorem on such a singular perturbation problem.

3. Proof of Theorem 2.1. To prove the existence of a minimizer, we define a comparison map as follows:

$$U_\epsilon(z) = \prod_{j=1}^{n} U_0\left(\frac{z - b_j}{\epsilon}\right) e^{i\, H_\epsilon(z)}, \tag{3.1}$$

for $z \in \Omega \subset \mathbb{C}$, where b_j's are n distinct points in Ω and H_ϵ is a real-valued smooth function in Ω such that

$$U_\epsilon = g \quad \text{on } \partial\Omega, \quad \|H_\epsilon\|_{C^2(\Omega)} = O(1).$$

Hereafter, U_0 is the symmetric vortex solution (cf. [4], [10], [11]) defined by

$$U_0(z) = f(R)\, e^{i\,\theta} \quad \text{for } z \in \mathbb{C}, \tag{3.2}$$

where $R = |z|$ and (R, θ) is the polar coordinate in $\mathbb{C}$. Moreover, $f(R)$ satisfies

$$\begin{cases} f'' + \frac{1}{R} f' - \frac{1}{R^2} f + (1 - f^2) f = 0 & \text{for } R > 0, \\ f(0) = 0, f(\infty) = 1. \end{cases} \tag{3.3}$$

From [4] and [11], the symmetric vortex solution U_0 satisfies the following lemma.

Lemma 3.1.
(i) $f(R) = \alpha_0 R + \alpha_1 R^3 + O(R^5)$ *as* $R \to 0+$, *where* $\alpha_0 > 0, \alpha_1 \in \mathbb{R}$ *are constants,*
(ii) $f(R) = 1 - \frac{1}{2R^2} + O(R^{-4})$ *as* $R \to +\infty$,
(iii) $U_0 = f(R)\, e^{i\theta}$ *is analytic in* $\mathbb{C}$.

Hence it is easy to check that

$$\hat{G}_\epsilon(U_\epsilon) = 2\pi n \log \frac{1}{\epsilon} + O(1) \quad \text{as } \epsilon \to 0+. \tag{3.4}$$

Now fix $0 < \epsilon \ll 1$. We claim that $\inf_{u \in H^1_g(\Omega)} \hat{G}_\epsilon(u)$ attains a minimizer $u_\epsilon \in H^2(\Omega)$. Let $\{u_k\}$ be a minimizing sequence such that

$$\hat{G}_\epsilon(u_k) \to \inf_{u \in H^1_g(\Omega)} \hat{G}_\epsilon(u). \tag{3.5}$$

Then by (2.2), (3.4), and (3.5), we have

$$\liminf_{k \to \infty} \int_\Omega |\nabla u_k|^2 + |\mathbb{B} u_k|^2 + |\Delta u_k|^2 \, dx\, dy < +\infty.$$

Hence there exists a subsequence $\{u_{k_j}\}$ such that

$$\|u_{k_j}\|_{H^2} \le K_\epsilon \quad \forall j \ge 1\,, \tag{3.6}$$

where $K_\epsilon > 0$ is a constant independent of j. Thus (3.6) implies

$$u_{k_j} \to u_\epsilon \quad \text{weakly in} \quad H^2(\Omega) \quad \text{as} \ \ j \to \infty\,. \tag{3.7}$$

Therefore by Fatou's lemma, u_ϵ is a minimizer of $\hat{G}_\epsilon$ over $H^1_g(\Omega)$.

From (2.2), (2.4), (3.4), and u_ϵ is a minimizer of $\inf_{u\in H^1_g(\Omega)} \hat{G}_\epsilon(u)$, we obtain

$$E_\epsilon(u_\epsilon) \le \pi\, n\, \log\, \frac{1}{\epsilon} \,+\, O(1)\,. \tag{3.8}$$

Moreover, by (3.8) and [29], we have

$$E_\epsilon(u_\epsilon) = \pi\, n\, \log\, \frac{1}{\epsilon} \,+\, O(1)\,. \tag{3.9}$$

Hence (3.4) and (3.9) imply that

$$\delta_\epsilon \int_\Omega |\mathbb{B} u_\epsilon|^2\, dx\, dy = O(1) \tag{3.10}$$

and

$$\gamma_\epsilon \int_\Omega |\Delta u_\epsilon|^2\, dx\, dy = O(1)\,. \tag{3.11}$$

Thus we complete the proof of (ii).

By (3.9) and Propositions 1.1 and 1.2 in [16], we complete the proof of (i). Furthermore, we obtain that u_ϵ converges to u_* (up to a subsequence) strongly in $L^2(\Omega)$ and weakly in $H^1_{loc}(\Omega\backslash\{a_1,\dots,a_n\})$, where $a_1,\dots,a_n \in \Omega$, $u_*(z) = \prod_{j=1}^n \frac{z-a_j}{|z-a_j|}\, e^{i\,h(z)} \forall z \in \Omega \subset \mathbb{C}$ and h is a real-valued function. Now we show that h is a harmonic function as follows: Consider the Euler–Lagrange equation of $\hat{G}_\epsilon$ with respect to the minimizer u_ϵ. Then u_ϵ satisfies

$$\Delta u_\epsilon + \frac{1}{\epsilon^2}(1-|u_\epsilon|^2)u_\epsilon - \delta_\epsilon\, \mathbb{B}^2 u_\epsilon - \gamma_\epsilon\, \Delta^2 u_\epsilon = 0 \quad \text{in } \Omega\,. \tag{3.12}$$

Perform the wedge product with u_ϵ and (3.12). This is a standard trick to erase the cubic nonlinear term in (3.12) (cf. [26] and [29]). Then we have

$$u_\epsilon \wedge \Delta u_\epsilon - \delta_\epsilon\, u_\epsilon \wedge \mathbb{B}^2 u_\epsilon - \gamma_\epsilon\, u_\epsilon \wedge \Delta^2 u_\epsilon = 0 \quad \text{in } \Omega\,. \tag{3.13}$$

Let $p \in C^\infty_0(\Omega)$ be a test function. Multiply (3.13) by p and integrate it on Ω. Then using integration by parts, we obtain

$$\begin{aligned} &-\int_\Omega (u_\epsilon \wedge \partial_x u_\epsilon) p_x + (u_\epsilon \wedge \partial_y u_\epsilon) p_y \\ &= \delta_\epsilon \int_\Omega (u_\epsilon \wedge \mathbb{B} u_\epsilon)\mathbb{B} p + 2(\partial_x u_\epsilon \wedge \mathbb{B} u_\epsilon) p_x - 2(\partial_y u_\epsilon \wedge \mathbb{B} u_\epsilon) p_y \\ &\quad + \gamma_\epsilon \int_\Omega (u_\epsilon \wedge \Delta u_\epsilon)\Delta p + 2(\partial_x u_\epsilon \wedge \Delta u_\epsilon) p_x + 2(\partial_y u_\epsilon \wedge \Delta u_\epsilon) p_y. \end{aligned} \tag{3.14}$$

Here we have used the following formulas:

$$u \wedge \Delta u = \partial_x (u \wedge \partial_x u) + \partial_y (u \wedge \partial_y u),$$
$$u \wedge \mathbb{B}^2 u = \mathbb{B}(u \wedge \mathbb{B}u) - 2(u_x \wedge \mathbb{B}u_x - u_y \wedge \mathbb{B}u_y),$$
$$u \wedge \Delta^2 u = \Delta(u \wedge \Delta u) - 2(u_x \wedge \Delta u_x + u_y \wedge \Delta u_y).$$

Hence by $0 < \gamma_\epsilon, \delta_\epsilon = O(\epsilon^2)$, (3.9)–(3.11), (3.14), and the Holder inequality, the limit map u_* satisfies

$$u_* \wedge \Delta u_* = 0 \quad \text{in distribution sense}. \tag{3.15}$$

Thus u_* is a canonical harmonic map, i.e., h is a harmonic function. Therefore we complete the proof of (iii).

Now we prove (iv) as follows: Let $(\tilde{a}_1, \ldots, \tilde{a}_n) \in \Omega^n$ be a global minimizer of the renormalized energy W_g. The definition of W_g can be found in [2]. Then we define another comparison map as follows:

$$V_\epsilon(z) = \begin{cases} u_\epsilon(z - \tilde{a}_j + a_j) & \text{if } z \in B_{\epsilon^\alpha}(\tilde{a}_j), j = 1, \ldots, n, \\ \tilde{U}_\epsilon(z) & \text{if } z \in \Omega_{\epsilon^\alpha} \equiv \Omega \backslash \cup_{j=1}^n B_{\epsilon^\alpha}(\tilde{a}_j), \end{cases} \tag{3.16}$$

where $0 < \alpha < 1$ is a constant and $\tilde{U}_\epsilon$ is a minimizer of E_ϵ over $H^1_{\tilde{g}}(\Omega_{\epsilon^\alpha})$. Here the boundary condition $\tilde{g}$ is defined by

$$\tilde{g} = \begin{cases} g & \text{on } \partial\Omega, \\ u_\epsilon(\cdot - \tilde{a}_j + a_j) & \text{on } \partial B_{\epsilon^\alpha}(\tilde{a}_j), j = 1, \ldots, n. \end{cases} \tag{3.17}$$

Hence by (iii), [2], and [29], $\tilde{U}_\epsilon$ satisfies

$$\tilde{U}_\epsilon \to \prod_{j=1}^n \frac{z - \tilde{a}_j}{|z - \tilde{a}_j|} e^{i\tilde{h}(z)} \quad \text{in } C^2(\Omega_{\epsilon^\alpha}) \quad \text{as } \epsilon \to 0+, \tag{3.18}$$

where $\tilde{h}$ is a harmonic function. The convergence of (3.18) may be up to a subsequence. However, this does not affect the following argument. Thus by (3.18) and [2], it is easy to check that

$$\hat{G}_\epsilon(V_\epsilon) = \sum_{j=1}^n \int_{B_{\epsilon^\alpha}(a_j)} \hat{g}_\epsilon(u_\epsilon) + 2\pi n\alpha \log \frac{1}{\epsilon} + 2W_g(\tilde{a}_1, \ldots, \tilde{a}_n) + o_\epsilon(1), \tag{3.19}$$

where $\hat{g}_\epsilon(u) = |\nabla u|^2 + \frac{1}{2\epsilon^2}(1 - |u|^2)^2 + \delta_\epsilon |\mathbb{B}u|^2 + \gamma_\epsilon |\Delta u|^2$ is the energy density of $\hat{G}_\epsilon$ and $o_\epsilon(1)$ is a small quantity which tends to zero as $\epsilon \to 0+$. On the other hand, by (iii) and [2], we have

$$\int_{\hat{\Omega}_{\epsilon^\alpha}} \frac{1}{2}|\nabla u_\epsilon|^2 + \frac{1}{4\,\epsilon^2}(1 - |u_\epsilon|^2)^2 \geq \pi n\alpha \log \frac{1}{\epsilon} + W_g(a_1, \ldots, a_n) + o_\epsilon(1), \tag{3.20}$$

where $\hat{\Omega}_{\epsilon^\alpha} = \Omega \backslash \cup_{j=1}^n B_{\epsilon^\alpha}(a_j)$. Hence (3.20) implies that

$$\hat{G}_\epsilon(u_\epsilon) \geq \sum_{j=1}^n \int_{B_{\epsilon^\alpha}(a_j)} \hat{g}_\epsilon(u_\epsilon) + 2\pi n\alpha \log \frac{1}{\epsilon} + 2W_g(a_1, \ldots, a_n) + o_\epsilon(1). \tag{3.21}$$

Thus by (3.19) and (3.21), we obtain

$$W_g(a_1, \ldots, a_n) \leq W_g(\tilde{a}_1, \ldots, \tilde{a}_n) + o_\epsilon(1). \tag{3.22}$$

Since $(\tilde{a}_1, \ldots, \tilde{a}_n)$ is a global minimizer of W_g, then we complete the proof of (iv) by (3.22).

4. Single vortex structure in the vortex core. In this section, we assume that the single vortex structure is in the vortex core B_{R_0}, where $R_0 > 0$ is a large constant satisfying $1 \ll R_0 \leq \frac{1}{\epsilon}$. Hereafter, we denote B_{R_0} as a disk in $\mathbb{R}^2$ with radius R_0 and center at the origin. To study the vortex structure in the vortex core, we restrict (1.9) in the vortex core B_{R_0} as follows:

$$\hat{G}(\Psi_d) = \int_{B_{R_0}} |\nabla \Psi_d|^2 + \frac{1}{2}(1 - |\Psi_d|^2)^2 + \beta\, |\mathbb{B}\Psi_d|^2 + \gamma\, |\Delta \Psi_d|^2 \, dx\, dy\,, \tag{4.1}$$

where $\gamma = C\beta, C > 0$ is a constant independent of β, and $\beta > 0$ is a small parameter tending to zero as $T \to T_c$. We investigate (4.1) with $\beta > 0$ a small parameter to see the phase transition of d-wave superconductors.

The Euler–Lagrange equation of (4.1) is

$$\Delta \Psi_d + (1 - |\Psi_d|^2)\Psi_d - \beta\, (\mathbb{B}^2 + C\, \Delta^2)\Psi_d = 0 \quad \text{in } B_{R_0}\,. \tag{4.2}$$

Note that $E \equiv \mathbb{B}^2 + C\, \Delta^2$ is an elliptic operator as $C > 0$. Moreover, by the Lax–Milgram theorem, $E : H_0^2(B_{R_0}\,; \mathbb{C}) \to H^{-2}(B_{R_0}\,; \mathbb{C})$ is invertible and we denote E^{-1} as its inverse. Hence the standard elliptic regularity theorem (cf. [20]) can be applied in (4.2).

We state the main result on (4.2) as follows.

THEOREM 4.1. *There exists a solution Ψ_d of* (4.2) *satisfying*

$$\Psi_d(z, \beta) = f(R)\, e^{i\,\theta} + \beta(a(R)\, e^{-4i\,\theta} + b(R)\, e^{4i\,\theta} + c(R))\, e^{i\,\theta} + O(\beta^2) \quad \text{as } \beta \to 0\,, \tag{4.3}$$

where a, b, and c are smooth real-valued functions.

Equation (4.3) implies that the d-wave order parameter Ψ_d is fourfold symmetric in the vortex core. In [27], we learn a well-approximated solution of (4.2) with fourfold symmetry. Here we find an exact solution of (4.2) with the fourfold symmetry.

Proof. Proof of Theorem 4.1.

To solve (4.2), we set

$$\Psi_d(z, \beta) = U_0(z) + \beta w_1(z) + \beta^2 w_2(z) + \beta^3 w(z, \beta)\,, \tag{4.4}$$

where U_0 is the symmetric vortex solution defined in (3.2) and (3.3). Here w_1 satisfies

$$L\, w_1 - E\, U_0 = 0 \text{ in } B_{R_0}\,, \quad w_1 = 0 \text{ on } \partial B_{R_0}\,, \tag{4.5}$$

where $Lv = \Delta v + (1 - |U_0|^2)v - 2(U_0 \cdot v)U_0$ is the linearized operator of (4.2) with respect to a trivial solution $(\Psi_d, \beta) = (U_0, 0)$. In addition, w_2 satisfies that

$$\begin{aligned} L\, w_2 &= 2(U_0 \cdot w_1)w_1 + |w_1|^2 U_0 + E\, w_1 \quad \text{in } B_{R_0}\,, \\ w_2 &= 0 \quad \text{on } \partial B_{R_0}\,. \end{aligned} \tag{4.6}$$

It is easy to check that

$$EU_0 = h_{-3}(R)\, e^{-3i\,\theta} + h_1(R)\, e^{i\,\theta} + h_5(R)\, e^{5i\,\theta}\,, \tag{4.7}$$

where h_{-3}, h_1, and h_5 are real-valued smooth functions. By [14], [18], [19], and [23], L is a bijection from $H_0^1(B_{R_0}\,; \mathbb{C})$ onto $H^{-1}(B_{R_0}\,; \mathbb{C})$. Hence by (4.5)–(4.7), we have

$$w_1 = a(R)\, e^{-3i\,\theta} + b(R)\, e^{5i\,\theta} + c(R)\, e^{i\,\theta}\,, \tag{4.8}$$

$$w_2 = \sum_{k=0}^{2} a_{1-4k}(R)\, e^{i\,(1-4k)\,\theta} \;+\; a_{1+4k}(R)\, e^{i\,(1+4k)\,\theta}\,, \tag{4.9}$$

where a, b, c, and $a_{1\pm 4k}$'s are smooth real-valued functions.

Taking (4.4) into (4.2), we obtain that

$$\begin{aligned} L\,w = \;& 2[(U_0\cdot(w_2+\beta w))w_1 + (U_0\cdot w_1)(w_2+\beta w)] + \beta|w_2+\beta w|^2 U_0 \\ & +\beta(U_0\cdot(w_2+\beta w))(w_2+\beta w) + 2(w_1\cdot(w_2+\beta w))U_0 \\ & +|w_1+\beta(w_2+\beta w)|^2(w_1+\beta(w_2+\beta w)) + Ew_2 + \beta E w \quad \text{in } B_{R_0}\,. \end{aligned} \tag{4.10}$$

Hence (4.10) is equivalent to

$$\begin{aligned} E^{-1}\,L\,w = \;& E^{-1}\{2[(U_0\cdot(w_2+\beta w))w_1 + (U_0\cdot w_1)(w_2+\beta w)] + \beta|w_2+\beta w|^2 U_0 \\ & +\beta(U_0\cdot(w_2+\beta w))(w_2+\beta w) + 2(w_1\cdot(w_2+\beta w))U_0 \\ & +|w_1+\beta(w_2+\beta w)|^2(w_1+\beta(w_2+\beta w))\} + w_2 + \beta w \quad \text{in } B_{R_0}\,. \end{aligned} \tag{4.11}$$

Note that (4.11) has a trivial solution $(w,\beta) = (w_3, 0)$, where w_3 satisfies that

$$\begin{aligned} L\,w_3 \;& = 2[(U_0\cdot w_2)w_1 + (U_0\cdot w_1)w_2] + 2(w_1\cdot w_2)U_0 \\ & \quad + |w_1|^2 w_1 + E\,w_2 \;\text{ in }\; B_{R_0}\,, \\ w_3 \;& = 0 \;\text{ on }\; \partial B_{R_0}\,. \end{aligned} \tag{4.12}$$

Since U_0, w_1, w_2 are smooth functions and L is bijective from $H_0^1(B_{R_0}\,;\mathbb{C})$ onto $H^{-1}(B_{R_0}\,;\mathbb{C})$, then the standard elliptic regularity theorem implies that w_3 is also a smooth function. Furthermore, since E is bijective from $H_0^2(B_{R_0}\,;\mathbb{C})$ onto $H^{-2}(B_{R_0}\,;\mathbb{C})$ and $H^{-1}(B_{R_0}\,;\mathbb{C})$ is embedded in $H^{-2}(B_{R_0}\,;\mathbb{C})$, then E is a bijection from $H_0^2(B_{R_0}\,;\mathbb{C})$ $\cap\, H^3(B_{R_0}\,;\mathbb{C})$ onto $H^{-1}(B_{R_0}\,;\mathbb{C})$. We denote E^{-1} as the inverse of E. Hence $E^{-1}\,L$ is a bijection from $H_0^1(B_{R_0}\,;\mathbb{C})$ onto $H_0^2(B_{R_0}\,;\mathbb{C}) \cap H^3(B_{R_0}\,;\mathbb{C})$. Thus by the implicit function theorem, (4.11) has a unique solution $w \in H_0^1(B_{R_0}\,;\mathbb{C})$ as $|\beta|$ is sufficiently small. Moreover, the standard elliptic regularity theorem may imply the smoothness of w. Therefore (4.2) has a solution Ψ_d satisfying (4.4) as $|\beta|$ is sufficiently small. By (4.4), (4.8), and (4.9), we obtain (4.3) and complete the proof of Theorem 4.1. ∎

Final remark. By (1.4) with $A \equiv 0$ and (4.3), we have

$$\Psi_s(z) = \xi\mathbb{B}[U_0 + \beta(a(R)\, e^{-4i\,\theta} + b(R)\, e^{4i\,\theta} + c(R))\, e^{i\,\theta} + O(\beta^2)] \tag{4.13}$$

as $\beta \to 0$. Since $U_0(z) = f(R)\, e^{i\,\theta}$, then

$$\mathbb{B}U_0(z) = \frac{1}{2}\left(f' + \frac{1}{R}\,f\right)'\, e^{-i\,\theta} + \frac{1}{2}\left[\left(f' - \frac{1}{R}f\right)' - \frac{2}{R}\left(f' - \frac{1}{R}\,f\right)\right] e^{3i\,\theta}\,. \tag{4.14}$$

Hence by (i), (ii) of Lemma 3.1 and (4.14), $\mathbb{B}U_0$ satisfies

$$\mathbb{B}U_0(z) = 4\alpha_1\, R\, e^{-i\,\theta} + O(R^3) \quad \text{as } R \to 0+ \tag{4.15}$$

and

$$\mathbb{B}U_0(z) = -\,\frac{1}{2\,R^2}\, e^{-i\,\theta} \;+\; \frac{3}{2\,R^2}\, e^{3i\,\theta} + O(R^{-4}) \quad \text{as } R \to +\infty\,. \tag{4.16}$$

By (4.15) and (4.16), the degree of $\mathbb{B}U_0$ is minus one in B_{r_1} and three in B_{R_1} as $0 < r_1 \ll 1$ and $R_1 \gg 1$. Moreover, by [4] and [11], it is easy to check that

$$\frac{d}{d\,z}\,\mathbb{B}\,U_0(z) \;\neq 0 \quad \text{if } \mathbb{B}U_0(z) = 0\,. \tag{4.17}$$

Hence (iii) of Lemma 3.1 and (4.17) imply that $\mathbb{B}U_0$ has only simple zeros in $\mathbb{C}$. Thus $\mathbb{B}U_0$ has a single zero with degree minus one at the origin and another four zeros with degree one away from the origin. Therefore as $|\beta|$ is sufficiently small, Ψ_s has a single zero with degree minus one near the origin and another four zeros with degree one away from the origin. This indicates the four-lobe structure of Ψ_s in the vortex core. The numerical simulation can be found in [7], [8], and [31].

Acknowledgments. The second author wishes to express his sincere thanks to B. Rosenstein for helpful discussions. He also sincerely thanks the referees for their suggestions.

REFERENCES

[1] I. Affleck, M. Franz, and M. H. S. Amin, *Generalized London free energy for high-T_c vortex lattices*, Phys. Rev. B, 55 (1997), pp. 704–707.
[2] F. Bethuel, H. Brezis, and F. Helein, *Ginzburg–Landau Vortices*, Birkhauser, Boston, 1994.
[3] A. J. Berlinsky, A. L. Fetter, M. Franz, C. Kallin, and P. I. Soininen, *Ginzburg–Landau theory of vortices in d-wave superconductors*, Phys. Rev. Lett., 75 (1995), pp. 2200–2203.
[4] X. Chen, C. M. Elliott, and T. Qi, *Shooting method for vortex solutions of a complex-valued Ginzburg–Landau equation*, Proc. Roy. Soc. Edinburgh Sect. A, 124 (1994), pp. 1075–1088.
[5] D. Chang, C.-Y. Mu, B. Rosenstein, and C. L. Wu, *The effect of anisotropy on vortex lattice structure and flux flow in d-wave superconductors*, Chinese J. Phys., 36 (1998), pp. 238–244.
[6] D. Chang, C.-Y. Mu, B. Rosenstein, and C. L. Wu, *Static and dynamical anisotropy effects in mixed state of d-wave superconductors*, Phys. Rev. B, 57 (1998), pp. 7955–7969.
[7] Q. Du, *Studies of a Ginzburg–Landau model for d-wave superconductors*, SIAM J. Appl. Math., 59 (1999), pp. 1225–1250.
[8] M. Franz, C. Kallin, P. I. Soininen, A. J. Berlinsky, and A. L. Fetter, *Vortex state in a d-wave superconductor*, Phys. Rev. B, 53 (1996), pp. 5795–5814.
[9] P. G. de Gennes, *Superconductivity of Metals and Alloys*, Addison-Wesley, Reading, MA, 1989.
[10] P. S. Hagan, *Spiral waves in reaction-diffusion equations*, SIAM J. Appl. Math., 42 (1982), pp. 762–786.
[11] R. M. Hervé, and M. Hervé, *Étude qualitative des solutions réelles d'une équation différentielle liée à l'équation de Ginzburg–Landau*, Ann. Inst. Henri Poincaré Anal. Non Linéaire, 11 (1994), pp. 427–440.
[12] J. D. Jorgensen, D. G. Hinks, O. Chmaissem, D. N. Argyriou, J. F. Mitchell, and B. Dabrowski, *Structural features that optimize high temperature superconductivity*, in Recent Developments in High Temperature Superconductivity, B. W. Veal, B. M. Dabrowski, and P. W. Klamut, eds., Springer, Berlin, Heidelberg, 1996, pp. 1–15.
[13] G. Kotliar, *Resonating valence bonds and d-wave superconductivity*, Phys. Rev. B, 37 (1988), pp. 3664–3666.
[14] E. H. Lieb, and M. Loss, *Symmetry of the Ginzburg–Landau minimizer in a disc*, Math. Res. Lett., 1 (1994), pp. 701–715.
[15] F. H. Lin, *Static and Moving Vortices in Ginzburg–Landau Theories*, Progr. Nonlinear Differential Equations Appl. 29, Birkhäuser, Basel, 1997.
[16] F. H. Lin, *Vortex dynamics for the nonlinear wave equation*, Comm. Pure Appl. Math., 52 (1999), pp. 737–761.
[17] F. H. Lin and T. C. Lin, *Minimax solutions of the Ginzburg–Landau equations*, Selecta. Math. (N.S.), 3 (1997), pp. 99–113.
[18] T. C. Lin, *The stability of the radial solution to the Ginzburg–Landau equation*, Comm. Partial Differential Equations, 22 (1997), pp. 619–632.
[19] P. Mironescu, *On the stability of radial solutions of the Ginzburg–Landau equation*, J. Funct. Anal., 130 (1995), pp. 334–344.
[20] C. B. Morrey, *Multiple Integrals in the Calculus of Variations*, Springer-Verlag, New York, 1966.
[21] K. A. Müller, *Possible coexistence of s- and d-wave condensates in copper oxide superconductors*, Nature, 377 (1995), pp. 133–135.

[22] K. Park and D. A. Huse, *The phase transition to a square vortex lattice in type*-II *superconductors with fourfold anisotropy*, Phys. Rev. B, 58 (1998), pp. 9427–9432.

[23] F. Pacard and T. Riviere, *Construction of Solutions of the Ginzburg–Landau Equation Having Regular Zero Set*, preprint.

[24] Y. Ren, J. H. Xu, and C. S. Ting, *Ginzburg–Landau equations and vortex structure of a* $d_{x^2-y^2}$ *superconductor*, Phys. Rev. Lett., 74 (1995), pp. 3680–3683.

[25] Y. Ren, J. H. Xu, and C. S. Ting, *Ginzburg–Landau equations for mixed* $s + d$ *symmetry superconductors*, Phys. Rev. B, 53 (1996), pp. 2249–2252.

[26] J. Shatah and M. Struwe, *Geometric Wave Equations*, Courant Lect. Notes Math. 2, New York University, New York, 1998.

[27] J. Shiraishi, M. Kohmoto, and K. Maki, *Vortex lattice transition in d-wave superconductors*, Phys. Rev. B, 59 (1999), pp. 4497–4503.

[28] P. I. Soininen, C. Kallin, and A. J. Berlinsky, *Structure of a vortex line in a* $d_{x^2-y^2}$ *superconductor*, Phys. Rev. B, 50 (1994), pp. 13883–13886.

[29] M. Struwe, *On the asymptotic behavior of minimizers of the Ginzburg–Landau model in* 2 *dimensions*, Differential Integral Equations, 7 (1994), pp. 1613–1624; erratum, 8 (1995), p. 224.

[30] Y. De Wilde, M. Iavarone, U. Welp, V. Metlushko, A. E. Koshelev, I. Aranson, G. W. Crabtree, and P. C. Canfield, *Scanning tunneling microscopy observation of a square Abrikosov lattice in* $LuNi_2B_2C$, Phys. Rev. Lett., 78 (1997), pp. 4273–4276.

[31] J. H. Xu, Y. Ren, and C. S. Ting, *Structures of single vortex and vortex lattice in a d-wave superconductor*, Phys. Rev. B, 53 (1996), pp. 2991–2994.

SIAM J. MATH. ANAL.
Vol. 32, No. 3, pp. 504–521

COEXISTENCE WITH CHEMOTAXIS*

LE DUNG†

Dedicated to Professor Jack Hale on his 70th birthday

Abstract. We study the existence of positive solutions to a coupled system of elliptic equations modeling competition in a bio-reactor with chemotactic response. A fixed point index technique is used to derive sufficient conditions for coexistence solutions.

Key words. chemotaxis, cross diffusion, fixed point index

AMS subject classifications. 35J55, 58J20

PII. S0036141099346779

1. Introduction.

The purpose of this paper is to consider the existence of positive solutions to the system

$$\begin{cases} -A_0 u_0 & = & -f_0(x,\vec{u}), & x \in \Omega, \\ -A_i u_i + \operatorname{div}(u_i \Phi_i(u_0)\nabla u_0) & = & f_i(x,\vec{u}), & x \in \Omega, \end{cases} \tag{1.1}$$

with the boundary conditions

$$\frac{\partial u_i}{\partial n} + r_i\left(x, \frac{\partial u_0}{\partial n}\right)u = u_0^i, \quad \frac{\partial u_0}{\partial n} + r_0(x)u_0 = S_0, \quad x \in \partial\Omega. \tag{1.2}$$

Here, $\vec{u} = (u_0, u_1, \ldots, u_m)$ and A_i are linear elliptic operators defined on an open bounded set Ω of $\mathbb{R}^n$. In this form (1.1) and (1.2) model the steady state solutions of a reaction-diffusion system describing the competition in a nonmixed bio-reactor where u_0 represents the concentration of a nutrient and u_i the densities of the species of cells (or bacteria). In addition to the derivative terms in the operator A_i reflecting the random diffusive flux and the convection effect in the model, the term $(u_i\Phi_i(u_0)(u_0)_x)_x$ reflects the chemotactic flux response of each species to the presence of the nutrient u_0. The function Φ_i, the so-called sensitivity rate, is included to indicate that the sensitivity of cells to the nutrient may vary at different levels of nutrient concentration.

Ever since the appearance of the work of Keller and Segel [17] on an aggregation model for the slime mold *Dictyostelium discoideum*, there has been great interest in modeling chemotaxis and in the mathematical analysis of systems like the Keller–Segel model. Here, we note the work of Schaff [22] and Lin, Ni, and Takagi [21] on steady states and the work of Jäger and Luckhaus [16] and Herrero and Velázquez [15] on finite time blowup of solutions.

*Received by the editors February 10, 1999; accepted for publication (in revised form) January 24, 2000; published electronically September 15, 2000.

http://www.siam.org/journals/sima/32-3/34677.html

†Georgia Institute of Technology, Center for Dynamical Systems and Nonlinear Studies, Atlanta, GA 30332. Current address: Division of Mathematics and Statistics, University of Texas at San Antonio, 6900 North Loop 1604 West, San Antonio, TX 78249 (dle@sphere.math.utsa.edu).

More recently, analytic work was done by Allegretto, Xie, and Yang [1] on a similar system of two equations modeling the chemotactic response of endothelial cells to the tumor angiogenesis factor. In [23], a bifurcation analysis was carried out to study the coexistence in unmixed chemostats with chemotactic effects following the pioneering work in this area by Lauffenburger and coworkers [18, 19, 20]. These authors assumed that the space dimension is one and used numerical methods to study the following system:

$$\begin{array}{lll} -S_{xx} & = & -f_1(S)u_1 - f_2(S)u_2, \\ -d_i(u_i)_{xx} + d_i[\Phi_i(S)u_i S_x]_x & = & [f_i(S) - k_i]u_i \end{array} \tag{1.3}$$

with boundary conditions

$$\begin{aligned} &\frac{\partial S}{\partial x}(0,t) = 0, \quad S(1,t) = 1, \\ &\frac{\partial u_i}{\partial x}(0,t) = 0 = \frac{\partial u_i}{\partial x}(1,t) - u_i(1,t)\Phi_i(S(1,t))\frac{\partial S}{\partial x}(1,t). \end{aligned}$$

The authors assume that the chemotactic sensitivity Φ follows the receptor law

$$\Phi_i(S) = \frac{\alpha_i}{(a_i + S)^2},$$

where α_i, a_i are positive constants. Many different forms have been used in the literature including constant $\Phi = \alpha$ and the log law, $\Phi = \alpha/(a + S)$. The functions $f_i(S)$ represent the functional response of the ith organism to nutrient concentration S and typically are bounded functions satisfying $f_i(0) = 0$, $f_i' > 0$. The constants k_i are cell death rates. One can see that the system (1.1), (1.2) (with $u_0 = S$) includes the above as its special case.

Recently, Wang [23] made use of bifurcation techniques to study the existence of positive steady states of a single species and one-dimensional system (1.3) and also obtained a result on global existence of time-dependent solutions. The argument seems to work only if the "washout state" (see section 3) is spatial homogeneous.

There has been great interest in finding conditions for the existence of at least one steady state of reaction-diffusion systems. Many authors have investigated this problem within the different contexts of biology, ecology, etc., using various techniques. Since the problem is nonvariational, one successful method is the use of topological index theory of fixed points applying to the fixed point form of the system in certain appropriate Banach spaces. We mention here the early works [2, 14]. The main difficulty is to establish certain a priori estimates for the solutions. The couplings in the diffusion terms make this task even more nontrivial. One should recall that there are examples, even with simple reaction terms, where the corresponding time-dependent system can exhibit blowup phenomena (see [15, 16]). Under reasonable conditions, we will obtain in section 2 a priori estimates for the L^∞ norm of solutions to (1.1). Once this is done, the existence of at least one solution follows from the Brouwer–Leray–Schauder degree theory in a standard fashion.

However, it can happen that certain solutions to the system can be found directly by simple inspections. For instance, the system (1.3) has homogeneous boundary conditions for u_i and therefore has trivially a solution with $u_i \equiv 0$ for $i \geq 1$ and appropriate S. Biologically speaking, this solution represents the situation where all the species are washed out from the culture; we then refer to this solution as the "washout" or trivial solution. In this case; the above existence conclusion does not

seem to give interesting results. The main purpose of this paper is to find sufficient conditions for other nontrivial solutions to exist.

Recently, efforts have been made to determine sufficient conditions for nontrivial or positive solutions of (1.4) as certain special (or trivial) solutions were already known. In particular, in [3, 4, 6, 9, 11] homotopy arguments have been employed to give sufficient conditions in terms of eigenvalue problems. The problem was usually considered without chemotactic effect and was equivalent to solving a system of elliptic equations

$$A_i u_i = f_i(x, u_0, \dots, u_m), \quad x \in \Omega,\ i = 0, \dots, m, \tag{1.4}$$

equipped with boundary conditions on $\partial\Omega$. The fixed point form then reads

$$U = K \circ F(U). \tag{1.5}$$

Here, K is the inverse of $\operatorname{diag}\{A_i\}$ and F is defined by the f_i's. The techniques in these works seem to be applicable only for problems of the form (1.4) which has the couplings that occur *only* in the reaction terms f_i. That is, the left-hand side of (1.4) depends only on u_i. Therefore, the methods in the aforementioned works are not well suited for strongly coupled elliptic systems whose fixed point form is not (1.5). On the other hand, one may try to write the fixed point form as $U = \Sigma(U)$ and make use of the degree theory in positive cones (see [2]). However, in doing so, one would quickly find that certain differentiability of the operator Σ needs to be verified. In our case, for this purpose alone, more estimates and extra smoothness assumptions would need to be made on the system (note also the couplings and nonlinearities in the boundary conditions (1.2)).

Nevertheless, in a joint work with Hal Smith [10], we were able to employ arguments similar to [6, 11] to give sufficient conditions for positive coexistence. However, certain unnecessary and biologically unrealistic restrictions on the sensitivity rate Φ_i and the space dimension n (≤ 3) had to be made.

In section 3 of this paper, we will generalize the techniques and revisit the results in [10]. Here, we are able to consider (1.1) in arbitrary dimension space and greatly relax the restriction on the functions Φ_i. Furthermore, we will present an abstract fixed point result for (1.1), (1.2). The existence results in section 2 are just simple consequences of the abstract ones. In fact, section 3.1 unifies the treatments in [3, 6, 11] and [10]. Moreover, they can also be used to study the existence of periodic solutions for strongly coupled parabolic systems (see [5]).

2. The model and existence results. Let Ω be a bounded domain in $\mathbb{R}^n$ with smooth boundary $\partial\Omega$. Consider the following elliptic operators

$$A_i u = \Delta u - b_i(x)\nabla u - c_i u, \quad i = 0, \dots, m,$$

and the elliptic system

$$\begin{cases} -A_0 u_0 = -f_0(x, \vec{u}), & x \in \Omega, \\ -A_i u_i + \operatorname{div}(u_i \Phi_i(u_0)\nabla u_0) = f_i(x, \vec{u}), & x \in \Omega, \\ \dfrac{\partial u_i}{\partial n} + r_i\left(x, \dfrac{\partial u_0}{\partial n}\right)u = u_0^i, \quad \dfrac{\partial u_0}{\partial n} + r_0(x)u_0 = S_0, & x \in \partial\Omega, \end{cases} \tag{2.1}$$

where $\vec{u} = (u_0, u_1, \dots, u_m)$. Hereafter, b_i are continuous vector-valued functions and c_i are continuous functions on Ω. The boundary data $u_0^i(x), S_0(x)$ are bounded functions on $\partial\Omega$.

In this section, we will consider nonhomogeneous boundary conditions in (2.1) and give sufficient conditions which guarantee certain a priori estimates for the solutions. The existence of at least one solution then follows from the standard Leray–Schauder theory.

We make the following assumptions on the characteristics of the system (2.1).

(H.1) For $i = 0, \dots, m$, $b_i \in C(\bar{\Omega}, \mathbb{R}^n)$, $c_i \in C(\bar{\Omega}, \mathbb{R})$. Also, we assume that $\Phi_i : \mathbb{R}_+ \to \mathbb{R}_+$ is continuous differentiable. Moreover, there exist C^2 functions $B_i(x)$, $i = 0, \dots, m$, such that $b_i(x) = \nabla B_i(x)$.

(H.2) $r_0 \in C(\partial\Omega, \mathbb{R}_+)$ and $r_0(x), c_0(x) \geq 0$. Moreover, $c_0(x) > 0$ or r_0 is not identically zero on $\partial\Omega$. $u_0^i(x), S_0(x)$ are nonnegative bounded continuous functions on $\partial\Omega$. For $i \geq 1$, we assume that $|r_i(x, \bullet)|$ is locally bounded.

(H.3) $f_0(x, \vec{u}) \geq 0$ for positive $\vec{u} = (u_0, \dots, u_m) \in \mathbb{R}_+^{m+1}$. Furthermore, for $i > 0$, $f_i(x, \vec{u})$ satisfies the following asymptotically linear growth condition:

$$|f_i(x, \vec{u})| \leq C(u_0)|\vec{u}| + C(u_0), \quad \text{for } |\vec{u}| \text{ large}, \tag{2.2}$$

where $C(u_0)$ is bounded if $|u_0|$ is bounded.

For later use we denote by S_* the unique (by (H.2)) solution of the equation

$$\begin{cases} -\Delta S_* + b_0(x)\nabla S_* + c_0 S_* = 0, & x \in \Omega, \\ \dfrac{\partial S_*}{\partial n} + r_0(x) S_* = S_0(x), & x \in \partial\Omega. \end{cases} \tag{2.3}$$

Remark 2.1. In (H.1) we have assumed that the convection vectors b_i are potential vector fields. In fact, this assumption is not essential for the proof of the following theorems. However, without this assumption one would have to restrict to domains Ω in $\mathbb{R}^3$ (see Remark 2.4 after the proof of Theorem 2.1). We remark that the condition $\Phi' > 0$ had to be assumed in [10]. On the other hand, as a trivial generalization using Sobolev inequalities, the growth condition in (H.3) can be relaxed to certain subcritical growth conditions.

We first show that the L^∞ norm of positive solutions can be controlled by their L^1 norm.

THEOREM 2.1. *Assume* (H.1)–(H.3). *Let $\vec{u}$ be any nonnegative bounded solutions to* (2.1). *Then there exists a continuous function $K_1(\cdot)$ such that*

$$\|\vec{u}\|_\infty \leq K_1(\|\vec{u}\|_1). \tag{2.4}$$

Therefore, to obtain a priori L^∞ estimates we need only to control the L^1 norms of the solutions. There are ways to achieve this. Being inspired by the system (1.3) (see also [10]), we can consider the following assumptions on the convection and reaction terms f_i.

(F.1) There is a continuously differentiable extension of Φ_i to all of $\mathbb{R}$ and

$$r_i\left(x, \frac{\partial u_0}{\partial n}\right) + \Phi_i(u_0)\frac{\partial u_0}{\partial n} + \frac{\partial B_i}{\partial n} \geq 0$$

on $\partial\Omega$ for all (not necessarily nonnegative) function u_0 satisfying $u_0 \leq S_*$ in Ω, and $\frac{\partial u_0}{\partial n} = S_0 - r_0 u_0$.

(F.2) There exist positive constants h_i, β, C such that

$$-h_0 f_0(x, \vec{u}) + \sum_{i=1}^m h_i f_i(x, \vec{u}) \leq \beta \sum_{i=0}^m h_i u_i + C \tag{2.5}$$

for all $\vec{u} = (u_0, \dots, u_m) \in \mathbb{R}_+^{m+1}$. Moreover, there is a positive constant θ such that

$$c_i - \Delta B_i(x) - \beta \geq \theta \quad \text{for all } x \in \Omega, \quad i = 1, \dots, m.$$

We then have the following theorem.

THEOREM 2.2. *Under the conditions* (F1), (F.2), *there is a positive constant* K_3 *independent of* $\vec{u}$ *such that*

$$\|u_i\|_1 \leq K_2, \quad i = 1, \dots, m. \tag{2.6}$$

The proof is simple by integrating the equations over Ω and adding the results (see [10, Theorem 2]). Combining the above estimates we have the following.

COROLLARY 2.3. *Under the conditions* (H.1)–(H.3) *and* (F1), (F.2), *for any* $\nu \in (0,1)$, *if* $u_0^i \in C^\nu(\partial\Omega)$, *then there is a positive constant* K_3 *independent of* $\vec{u}$ *such that*

$$\|u_i\|_{C^\nu(\Omega)} \leq K_3, \quad i = 0, \dots, m. \tag{2.7}$$

Having established these a priori estimates, we study the solvability of (2.1) using Leray–Schauder degree theory. Since we will be interested only in nonnegative solutions and make use of maximum principles we will assume hereafter the following.

(F.3) f_i is continuous in its arguments and $f_i(x, \vec{u}) \geq 0$ for all $i > 0$, $x \in \Omega$, and $\vec{u} \in \mathbb{R}_+^m$. Moreover, $f_0(x, \vec{u})$ vanishes if $u_0 = 0$. Assume also that $u_0^i, S_0 \in C^\nu(\Omega)$ for some $\nu \in (0,1)$.

We should remark that the assumption $f_i(x, \vec{u}) \geq 0$ is not really needed here. In fact, because the solutions of (2.1) are shown to be uniformly bounded by Corollary 2.3, this condition can be fulfilled by adding ku_i, with k sufficiently large, to both sides of the ith equation and redefining f_i.

We set up a fixed point equation in the positive cone X_+ of $X := \prod_0^{m+1} C^\nu(\Omega)$ for (2.1) as follows (compare to [10, pp. 305–306]). Given $\psi = (\psi_0, \dots, \psi_m)$ in X_+, we denote by $S, u_1, \dots, u_m$ the solutions to the system

$$\begin{cases} -A_0 S = -f_0(x, \psi), & x \in \Omega, \\ -A_i u_i + \operatorname{div}(u_i \Phi_i(S) \nabla(S)) = f_i(x, \psi), & x \in \Omega, \\ \dfrac{\partial u_i}{\partial n} + r_i\left(x, \dfrac{\partial S}{\partial n}\right) u_i = u_0^i, \quad \dfrac{\partial S}{\partial n} + r_0(x) S = S_0, & x \in \partial\Omega. \end{cases} \tag{2.8}$$

We then define $F(\psi) = \vec{u} = (u_0, \dots, u_m)$ with $u_0 = [S]_+$ and u_i, for $i > 0$, as above. We have the following lemma.

LEMMA 2.2. *Assume that* (H.1)–(H.3) *and* (F.1)–(F.3) *hold. Then* $F : X_+ \to X_+$ *is a well-defined completely continuous operator. Moreover, fixed points of* F *in* X_+ *are solutions of* (2.1).

Next, we have the following consequence of the a priori estimates derived in Theorem 2.1.

LEMMA 2.3. *Let the assumptions of Lemma* 2.2 *hold. Then there is an* $R > 0$ *such that*

$$F(\lambda U) = U, \quad \lambda \in [0,1], \tag{2.9}$$

has no solution $U \in X_+$ *satisfying* $\|U\| = R$.

The above result allows us to compute the index of the map F and then obtain the existence of at least one solution to the system as follows.

THEOREM 2.4. *For $r > 0$, let $C_r = \{u \in X_+ : \|u\| < r\}$. With R given by* Lemma 2.3*, we have*

$$\operatorname{ind}(F, C_R) = +1.$$

In particular, there is a fixed point of F in C_R.

Proof. The fact that $\operatorname{ind}(F, P_R) = +1$ follows by using a standard homotopy using the parameter λ in (2.9) and (2.16). In particular, we consider the family of maps $H(\lambda, U) = F(\lambda U)$, which is a well-defined homotopy in C_R thanks to Lemma 2.3, and find

$$\operatorname{ind}(F, C_R) = \operatorname{ind}(H(0, \bullet), C_R) = +1.$$

The last equality follows from the fact that $H(0, U) = F(0)$ is a constant map (cf. Lemma 2.2). □

We conclude this section by giving the proofs of the stated lemmas and theorems.

2.1. A priori estimates. The fact that $f_0(x, \vec{u}) \geq 0$ and comparison principles give immediately the estimate $u_0 \leq S_*$. This and (H.1), (H.2), and the boundary condition for u_0 imply that there exists a positive constant $C(S_*)$ such that

$$|\Phi_i(u_0)|, \quad \left|\frac{\partial u_0}{\partial n}\right|, \quad \left|r_i\left(x, \frac{\partial u_0}{\partial n}\right)\right| \leq C(S_*) \quad \text{on } \partial\Omega. \tag{2.10}$$

We should remark that (2.10) does not give us any information on the derivative ∇u_0 inside the domain Ω.

Proof of Theorem 2.1. Dropping the subscripts, we write the equation for $u = u_i$ in the form

$$\begin{cases} -\operatorname{div}(\nabla u - (\nabla B + \Phi(u_0)\nabla u_0)u) + (c - \Delta B)u = f(x, \vec{u}) & \text{on } \Omega, \\ \dfrac{\partial u}{\partial n} + ru = u_0^i & \text{on } \partial\Omega. \end{cases} \tag{2.11}$$

Here we use the fact that $\nabla B \nabla u = \operatorname{div}(u\nabla B) - u\Delta B$. Set $\Phi_*(S) = \int_0^S \Phi(s)ds$. We introduce new variables

$$U = e^{-(B+\Phi_*(u_0))}u, \ \vec{U} = E(u_0)\vec{u}, \ \text{with } E(u_0) = \operatorname{diag}\{e^{-(B_i+\Phi_{*,i}(u_0))}\},$$

and observe the following:

$$\begin{aligned} \nabla u - (\nabla B + \Phi(u_0)\nabla u_0)u &= e^{B+\Phi_*(u_0)}\nabla U, \\ \frac{\partial u}{\partial n} &= e^{B+\Phi_*(u_0)}\left(\frac{\partial U}{\partial n} + \frac{\partial(B + \Phi_*(u_0))}{\partial n}U\right). \end{aligned}$$

We can see that (2.11) implies

$$\begin{cases} -\operatorname{div}(e^{B+\Phi_*(u_0)}\nabla U) + (c - \Delta B)e^{B+\Phi_*(u_0)}U = f(x, E^{-1}(u_0)\vec{U}) \text{ on } \Omega, \\ \dfrac{\partial U}{\partial n} + \left(r + \dfrac{\partial(B + \Phi_*(u_0))}{\partial n}\right)U = u_0^i e^{B+\Phi_*(u_0)} \text{ on } \partial\Omega. \end{cases} \tag{2.12}$$

Our system becomes a quasi-linear system of elliptic equations with the couplings occurring in the diffusion rates $e^{B+\Phi_*(u_0)}$. However, the new system is nondegenerate. That is, since $u_0 \geq 0$ and u_0 is bounded, we can find positive constants λ, Λ (depending on S_*) such that

$$0 < \lambda \leq e^{B+\Phi_*(u_0)} \leq \Lambda. \tag{2.13}$$

This also shows that the L^1 norms of U and u are comparable. Moreover, the growth condition (2.2) now reads

$$|f(x, E^{-1}(u_0)\vec{U})| \leq C(S_*)|\vec{U}| + C(S_*) \quad \text{for } |\vec{U}| \text{ large.} \tag{2.14}$$

Since solutions of the uniformly elliptic system (2.12) are also solutions of the corresponding parabolic system, whose nonlinearities satisfy (2.14), a more general result [8, 7] can be applied here to give the estimate (2.4) for $\vec{U}$. The same is true for $\vec{u}$ since the L^∞ norms of $\vec{U}, \vec{u}$ are comparable. The proof is complete. □

Remark 2.4. Without the assumption that b_i are potential vector fields, one can still make use of the change of variables $U = e^{-\Phi_*(u_0)}u$, but the equation for U in (2.12) will have the term $U\nabla u_0$. This will cause some difficulties in using the Moser-type iteration technique to obtain (2.4). One needs to control the term ∇u_0 but this quantity depends on some integral of u from the equation of u_0. Therefore, certain restrictions such as $n \leq 3$ should be made. We refer to [10, page 299] for the details. However, we want to remark that the assumption $\Phi' \geq 0$ in [10] is unnecessary due to the argument in this remark and by repeating the lines in [10, page 300]. On the other hand, we should remark that the above result does not hold for the parabolic-elliptic systems of Keller–Segel [15, 16, 17] where blowup may occur even though the L^1 norms of solutions are conserved. However, in the one-dimensional case, we are able to show in [5] that (2.4) continues to hold for parabolic versions of (2.1).

Proof of Corollary 2.3. From Theorems 2.1 and 2.2 we see that the L^∞ norms of the right-hand sides of the equations of (2.1) are bounded by some finite constant independent of u_i. From the L^p theory of elliptic equation (see [12, Theorem 19.1, p. 74]) we have

$$\|u_0\|_{W^{2,p}(\Omega)} \leq C_p(\|u_0\|_p + \|f_0(x, \vec{u})\|_p) \leq K(p)$$

for all $p \in (1, \infty)$. $C_p, K(p)$ are constants independent of S. Taking p sufficiently large and using the Sobolev imbedding theorem, we can see that $\|u_0\|_{C^{1+\nu}(\Omega)}$ is bounded. Thus, if we write the equation of u_i in its divergence form

$$-\mathrm{div}(\nabla u_i - \Phi_i(u_0)\nabla u_0 u_i) + b_i(x)\nabla u_i + c_i u = f_i(x, \vec{u}),$$

which has bounded Hölder continuous coeffients, then u_i, as a bounded weak solution to the above equation, is C^β Hölder continuous with bounded Hölder norm (see [13, Chapter 8]) for some $\beta \in (0, 1)$. Now, with $u_i \in C^\beta(\Omega)$ with uniformly bounded norms, we see that the right-hand sides of the equations are also bounded in $C^\beta(\Omega)$. The Schauder estimates (see [13, Chapter 6]) imply that $u_0 \in C^{2+\beta}(\Omega)$. This improves the regularity of the coefficients of the equation for u_i, $i > 0$, considering u_0 as a parameter. Applying Schauder's estimates again we can conclude that u_i belongs to $C^{2+\beta}(\Omega)$ as well. In addition, the $C^{2+\beta}$ norms of u_i are uniformly bounded. Our proof is complete. □

2.2. Existence.

Proof of Lemma 2.2. The proof is standard and we will only sketch the main points here (cf. [10, Lemma 3.1]). By (H.2), it is well known that the first equation has a unique solution u_0 in $C^{2+\nu}(\Omega)$ (see [13, Theorem 6.31]). Clearly, $u_0 = [S]_+ \geq 0$. By the maximum principle argument we can show that $S_* \geq S$. On the other hand, in the equation for $u = u_i$ we set $\Phi_*(S) = \int_0^S \Phi(s)ds$ and introduce the new variable $U = e^{-(B+\Phi_*(S))}u$ as in the proof of Theorem 2.1 to find that

$$\begin{cases} -\mathrm{div}(a(x)\nabla U) + d(x)U = g(x) & \text{on } \Omega, \\ \dfrac{\partial U}{\partial n} + R(x)U = 0 & \text{on } \partial\Omega \end{cases} \tag{2.15}$$

with

$$a(x) = e^{B+\Phi_*(S)}, \quad d(x) = (c - \Delta B)e^{B+\Phi_*(S)}, \quad g(x) = f(x, \vec{\psi}(x)),$$

$$R(x) = r\left(x, \frac{\partial S}{\partial n}(x)\right) + \frac{\partial B}{\partial n}(x) + \Phi(S(x))\frac{\partial S}{\partial n}(x).$$

The boundary condition of u_0 implies $\frac{\partial u_0}{\partial n} = S_0 - r_0 u_0$, so that, by (F.1), $R(x) \geq 0$ on $\partial\Omega$. Also, by (F.2), $d(x) > 0$. (In fact, this assumption in (F.2) is used only for the L^1 estimate of Theorem 2.2. Here, one can always achieve this by adding Ku_i to both sides of the equation, for K large enough.) The maximum principle for linear elliptic equations [13, Corollary 3.2] applies to the above equation for U ($g(x) \geq 0$) and shows that $U(x) > 0$. Therefore, $u(x) > 0$ in Ω.

Therefore, one can solve for S from its equation in (2.8) and substitute the result into the equation for U. The above facts about the coefficients of the equation of U also imply that we can solve for U_i, and thus u_i, uniquely in $C^{2+\nu}(\Omega)$ from their equations in (2.8). Hence, F is well defined and maps X_+ into itself. The complete continuity of F follows in a standard way using Schauder's estimates for uniformly elliptic equations (see Corollary 2.3). Finally, if $(u_0, \ldots, u_m) \in X_+$ is a fixed point of F then, because $f_0(x, u_0, \ldots, u_m) = 0$ as $u_0 = 0$ and (H.2), a maximum principle argument shows that $S \geq 0$ so that $u_0 \equiv S$. In this case, (2.8) and (2.1) coincide and our last claim follows. □

Proof of Lemma 2.3. The above equation (2.9) is equivalent to

$$\begin{cases} -A_0 u_0 & = & -f_0(x, \lambda\vec{u}), & x \in \Omega, \\ -A_i u_i + \mathrm{div}(u_i \Phi_i(u_0)\nabla u_0) & = & f_i(x, \lambda\vec{u}), & x \in \Omega, \\ \dfrac{\partial u_i}{\partial n} + r_i\left(x, \dfrac{\partial u_0}{\partial n}\right)u_i = 0, & & \dfrac{\partial u_0}{\partial n} + r_0(x)u_0 = S_0, & x \in \partial\Omega. \end{cases} \tag{2.16}$$

Define f_λ for $\lambda \in [0,1]$ by $f_\lambda = (\hat{f}_0, \ldots, \hat{f}_m)$, where $\hat{f}_0(x, \vec{u}) = f_0(x, \lambda\vec{u})$, $\hat{f}_i(x, \vec{u}) = f_i(x, \lambda\vec{u})$, $1 \leq i \leq m$. Then it is easy to check that if f satisfies (F.1) and (F.2) of section 2, which we are assuming, then f and f_λ, with $\lambda \in [0,1]$, also satisfy these assumptions with a common set of constants h_i, β, C, θ , which are independent of $\lambda \in [0,1]$. Therefore the L^1 estimates of Theorem 2.2 and the L^∞ estimates of Theorem 2.1 hold for the solutions of (2.16) using the same constants K_1, K_2. Consequently, we may take $R = K_3$, the constant given in Corollary 2.3, to complete our proof. □

3. Existence of positive steady states. As we mentioned in the introduction, in applying Theorem 2.4 to the system (1.3) we do not obtain interesting results, unless u_0^i is not identically zero, since we already observe that $u_0 = S_*$ and $u_i = 0$ for $i > 0$ is a solution of (1.3). This solution describes the situation when all species except the nutrient have been washed out from the reactor. In fact, this is a common scenario in biological models when certain "trivial" solutions have been known to exist by simple inspections. Hence, it is important to find conditions which guarantee the existence of other solutions which have positive components. Such solutions are called the *coexistence* states of the system.

In this section we will look at this question for the system (2.1) and restrict ourselves to the case $m \leq 2$. In biological terms, we will consider the coexistence problem for models of one nutrient and no more than two species. We then consider the following conditions.

(F.4) We assume that $u_0^i \equiv 0$ on $\partial\Omega$. For $i > 0$, $f_i(x, \vec{u})$ vanishes when $u_i = 0$ and $f_0(x, \vec{u}) = 0$ if $u_i = 0$ for all $i \neq 0$. Moreover, f_i has continuous partial derivatives $\frac{\partial f_i}{\partial u_i}$ satisfying $\frac{\partial f_i}{\partial u_i} > 0$.

Again, the assumption that $\frac{\partial f_i}{\partial u_i} > 0$ is not essential here since we can add ku_i, k sufficiently large, to both sides of the ith equation and redefine f_i.

It is clear that (2.1) has a solution $(S_*, 0, \ldots, 0)$ to which we refer as the "washout" solution of (2.1) and recall that $S_*(x) > 0$ for all x. From (F.4), we see that there can be two types of solutions to (2.1) in addition to this washout solution:

$$\text{Semitrivial solutions:} \quad (u_0, u_1, 0) \quad (u_0, 0, u_2); \quad \text{Positive solutions:} \quad (u_0, u_1, u_0),$$

where the components $u_i > 0$. In what follows we will study the existence of solutions of these types. To this end, we will make use of the abstract results presented below (section 3.1) to compute the fixed point index of the map F. To start, let us rewrite the fixed point form of (2.1) in the following way.

From the proof of Lemma 2.2, for any $\psi \in X_+$ given, we can write $u_0 = F_0(\psi)$ given by

$$u_0 = [S]_+, \text{with } -A_0 S = -f_0(x, \psi(x)) \quad \text{and} \quad \frac{\partial S}{\partial n} + r_0(x)S = S_0 \quad \text{on } \partial\Omega. \tag{3.1}$$

Substitute u_0, which depends on ψ, into the other equations. From the proof of Lemma 2.2, (2.15) is regularly elliptic for any such u_0. Thus we can define

$$K_i(\psi) = \left\{ -A_i \bullet + \alpha \operatorname{div}(\bullet \Phi_i(u_0) \nabla u_0), \frac{\partial \bullet}{\partial n} + r_i\left(x, \frac{\partial u_0}{\partial n}\right) \bullet \right\}^{-1}, \tag{3.2}$$

$$\bar{F}_i(\psi)(x) = f_i(x, \psi(x)).$$

Hence, $K_i(\psi)$ is a linear strongly positive and compact map for each $\psi \in X_+$. We can write $u = F(\psi)$ in the form

$$u_0 = F_0(\psi), \quad u_i = F_i(\psi) := K_i(\psi) \circ \bar{F}_i(\psi). \tag{3.3}$$

It is the purpose of the next section to study the fixed point problem for a system in which the operator is of the form (3.3). We want to remark that the result applies trivially to the nonchemotactic problems as well so that the following can be considered as an abstract version of the cases treated in [3, 6, 10, 11].

3.1. An index result. In order to state our main result of this section we will need some notation. Let X_i, $i = 0, \dots, m$, be ordered Banach spaces with positive cones C_i. Set $X = \oplus_{i=0}^m X_i$ and $C = \oplus_{i=0}^m C_i$.

Let Ω be an open set in C containing 0 and let $F_i : \bar{\Omega} \to C_i$ be completely continuous operators. Denote by $u = (u_0, \dots, u_m)$ a generic element in C. We define $F : \bar{\Omega} \to C$ by

$$F(u) = (F_0(u), \dots, F_m(u)).$$

Let $\mathcal{B}$ be the collection of subsets of $M = \{0, \dots, m\}$. For each nonempty $\beta \in \mathcal{B}$, we write $X_\beta = \oplus_{i\in\beta} X_i$, $C_\beta = \oplus_{i\in\beta} C_i$. If $\beta = \emptyset$ we set $X_\beta = C_\beta = \{0\}$. In what follows, we will regard X_β, C_β as subspaces and subsets of X. We also define

$$F_\beta(u) : \bar{\Omega} \to C_\beta, \quad F_\beta(u) = (F_i(u))_{i\in\beta}.$$

For each $\beta \in \mathcal{B}$, we set

$$Z_\beta = \{u = (u_0, \dots, u_m) \in \Omega \ : \ F(u) = u, \quad u_i > 0 \quad \text{if } i \in \beta, \quad u_i = 0 \quad \text{if otherwise}\}.$$

Thus, Z_β is the set of fixed points of F on the "face" (or edge) C_β of C. In particular, if $\beta = M$, then Z_β consists of all positive fixed points, that is, all of their components are positive. Otherwise, we refer to Z_β ($\beta \neq M$) as the set of semitrivial fixed points. Also, $Z_\emptyset = \{0\}$ if $F(0) = 0$. If $\beta \neq \emptyset$ we can also think of Z_β as the set of positive fixed points of $F_\beta|_{C_\beta}$.

If $Z_\beta \neq \emptyset$ we set $\alpha = M \setminus \beta$ and assume that there exist open sets $W_\beta \subset \Omega \cap C_\beta$ and $W_\alpha \subset \Omega \cap C_\alpha$ with $0 \in W_\alpha$ such that $Z_\beta \subset W_\beta$. For $u \in \Omega$, we write $u = (u_\beta, u_\alpha)$ and consider the neighborhood $W_\beta \oplus W_\alpha$ of Z_β in Ω. For each fixed $u_\beta \in W_\beta$, $u_\alpha \in W_\alpha$ we assume that $F_\alpha(u_\beta, u_\alpha)$ can be expressed in the form

$$F_\alpha(u_\beta, u_\alpha) = K_\alpha(u_\beta, u_\alpha) \circ \bar{F}_\alpha(u_\beta, u_\alpha), \tag{3.4}$$

where $K_\alpha : W_\beta \oplus W_\alpha \to L(X_\alpha)$, the Banach space of bounded linear maps from X_α into itself, and $\bar{F}_\alpha : W_\beta \oplus W_\alpha \to X_\alpha$. Assume also that K_α is continuous in (u_β, u_α) and $\bar{F}_\alpha(u_\beta, \cdot) : W_\alpha \to X_\alpha$ is continuously differentiable (with respect to u_α) with its partial derivative denoted by $\partial_\alpha \bar{F}_\alpha$. We assume that $\partial_\alpha \bar{F}_\alpha(u_\beta, u_\alpha)$ is continuous in $W_\beta \oplus W_\alpha$.

For $\beta \in \mathcal{B}$ such that $Z_\beta \neq \emptyset$, we assume that

$$\bar{F}_\alpha(u_\beta, 0) = 0 \quad \text{for all } u_\beta \in W_\beta. \tag{3.5}$$

Set $B_\alpha(u_\beta, u_\alpha) = K_\alpha(u_\beta, u_\alpha) \circ \partial_\alpha \bar{F}_\alpha(u_\beta, u_\alpha)$. We consider the following condition.

(E_1) For each $u = (u_\beta, 0) \in Z_\beta$, 1 is not an eigenvalue of $B_\alpha(u_\beta, 0)$ corresponding to a positive eigenvector of $B_\alpha(u_\beta, 0)$.

Remark 3.1. If F_α is differentiable, one can easily see that (3.5) implies that $B_\alpha(u_\beta, 0)$ coincides with $\partial_\alpha F_\alpha(u_\beta, 0)$. However, since we do not assume any differentiability on K_α, $\partial_\alpha F_\alpha(u_\beta, u_\alpha)$ is not differentiable in general. We keep such notation to emphasize this fact.

Consider the following situations for the spectral radius r of $B_\alpha(u_\beta, 0)$:

(E_+) $r(B_\alpha(u_\beta, 0)) < 1$ for all $u_\beta \in Z_\beta$;

(E_-) $r(B_\alpha(u_\beta, 0)) > 1$ for all $u_\beta \in Z_\beta$.

We then set

$$\sigma(\beta) = \begin{cases} 1 & \text{if } (E_+) \text{ holds,} \\ 0 & \text{if } (E_-) \text{ holds.} \end{cases} \tag{3.6}$$

For each $\beta \in \mathcal{B}$ such that $\beta \neq \emptyset$ and $Z_\beta \neq \emptyset$, there exists an open neighborhood U_β of Z_β, as a subset of C_β, such that the set of fixed points of F_β in $\Omega \cap C_\beta$ is exactly Z_β. So, $\mathrm{ind}(F_\beta|_{C_\beta}, U_\beta)$ is well defined. If 0 is a fixed point of F, then $Z_\emptyset = \{0\}$ and, by Lemma 3.4 below, (E_1) implies that such a neighborhood $U_\emptyset$ of 0 in C exists. We then define the "face" indices

$$i(\beta) = \begin{cases} \mathrm{ind}(F_\beta|_{C_\beta}, U_\beta), & \beta \neq \emptyset, \\ \mathrm{ind}(F, U_\emptyset), & \beta = \emptyset. \end{cases} \tag{3.7}$$

If U is an open set and Z is the set of fixed points of F in U, then we also define the local index of the fixed point set Z by $i(F, Z) = \mathrm{ind}(F, U)$.

The index result then reads as follows.

THEOREM 3.1. *Suppose that* (3.5), (E_1), *and either* (E_+) *or* (E_-) *hold for each* β. *Moreover,* $B_\alpha(u_\beta, 0)$ *is strongly positive for all* $u_\beta \in Z_\beta$. *We assert that*

(i) $i(F, Z_\beta) = \sigma(\beta) i(\beta)$. *Here,* $i(F, Z_\beta)$ *is the local index of the fixed point set* Z_β;

(ii) *if*

$$\mathrm{ind}(F, \Omega) \neq \sum_{\beta \in \mathcal{B}, \beta \neq M, Z_\beta \neq \emptyset} \sigma(\beta) i(\beta), \tag{3.8}$$

then there exists at least a positive solution to $F(u) = u$.

Remark 3.2. The assumption that $B_\alpha(u_\beta, 0)$ is strongly positive can be dropped. The proof of Proposition 3.4 reveals that instead of (E_+), (E_-), we need only to assume that

(E'_+) For $u_\beta \in Z_\beta$, $B_\alpha(u_\beta, 0)$ has no positive eigenvector corresponding to an eigenvalue greater than 1.

(E'_-) For each $\beta \in \mathcal{B}$ with $Z_\beta \neq \emptyset$, there exists a positive $p_\alpha \in C_\alpha$ such that $v = B_\alpha(u_\beta, 0)v + tp_\alpha$ has no positive solution v for all $t > 0$.

Regarding (E'_-), we remark that it is easy to verify, especially when Z_β is a singleton (see Remark 3.5). In this case, we need only to assume that $B_\alpha(u_\beta, 0)$ has a positive eigenvector p_α corresponding to an eigenvalue $\lambda_\alpha > 1$.

Remark 3.3. In terms of stability of fixed points, condition (E_+) (resp., (E_-)) simply says that every element of Z_β is "stable" (resp., "unstable") to invasion by the α-variable ($\alpha = M \setminus \beta$). We also say that Z_β is "attracting" (resp., "repelling") in the complementary direction C_α of the cone C_β. We then observe that the right-hand side of (3.8) is just the sum of the face indices of attracting semitrivial fixed point set Z_β.

As a simple consequence of Theorem 3.1, we assert the following.

COROLLARY 3.2. *If* $\mathrm{ind}(F, \Omega) \neq 0$ *and* Z_β *is "repelling" for all* β *such that* $Z_\beta \neq \emptyset$, *then there is at least one positive fixed point of* F *in* Ω.

To prove Theorem 3.1 we need to compute the index of the map in a neighborhood of the "face" fixed point set Z_β. To this end, we need the following lemmas.

Suppose X_1 and X_2 are ordered Banach spaces with positive cones C_1 and C_2, respectively. We define $X = X_1 \oplus X_2$ and $C = C_1 \oplus C_2$.

Let Ω be an open set in C containing 0 and let $F_i : \bar{\Omega} \to C_i$ be completely continuous operators, $i = 1, 2$. Denote generic elements in C by (u, v) with $u \in C_1$ and $v \in C_2$. Let $F : \bar{\Omega} \to C$, a nonlinear map given by

$$F(u, v) = (F_1(u, v), F_2(u, v)). \tag{3.9}$$

We consider the following fixed point problem.

$$F(u, v) = (u, v). \tag{3.10}$$

Suppose $U \subset C_1$, $V \subset C_2$ are bounded open sets and

$$\bar{F}_2(u, 0) = 0 \quad \text{for all } u \in \bar{U}, \tag{3.11}$$

$$F_1(u, 0) \neq u \quad \text{for all } u \in \partial U. \tag{3.12}$$

Suppose that $Z_1 = \{u \in U \ : \ u = F_1(u, 0)\}$, the set of fixed points of $F_1(u, 0)$ in C_1, is not empty. By (3.11), if $u \in Z_1$ then $(u, 0)$ is a solution of (3.10) on the face C_1. As in (3.4), for each $u \in U$, we assume that the operator F_2 can be expressed in the form

$$F_2(u, v) = K(u, v) \circ \bar{F}_2(u, v), \tag{3.13}$$

where $K : U \oplus V \to L(X_2)$, the Banach space of bounded linear maps from X_2 into itself, and $\bar{F}_2 : U \oplus V \to X_2$. Assume also that K is continuous in u, v (not necessarily differentiable) and $\bar{F}_2(u, \cdot) : V \to X_2$ is continuously differentiable in v with its partial derivative denoted by $\partial_v \bar{F}_2(u, v)$. We suppose that $\partial_v \bar{F}_2(u, v)$ is continuous in $U \oplus V$ and denote

$$B(u, v) = K(u, v) \circ \partial_v \bar{F}_2(u, v), \quad (u, v) \in U \oplus V. \tag{3.14}$$

Then $B(u, v)$ is a map from $U \oplus V$ into C_2. Consider the following condition.

(B) For each $u \in Z_1$, $B(u, 0)$ is compact and 1 is not an eigenvalue of $B(u, 0)$ corresponding to a positive eigenvector.

The following decoupling technique is useful in computing the index of $F(u, v)$.

PROPOSITION 3.3. *Assume* (B) *and that* $tv \in V$ *for any* $v \in V$ *and* $t \in [0, 1]$. *Suppose that*

$$F(u, v) = (u, v) \quad \textit{has no solution in closure } (U \oplus V) \textit{ with } v > 0. \tag{3.15}$$

Then $\operatorname{ind}(F, U \oplus V) = \operatorname{ind}(H, U \oplus V)$, *where* $H(u, v) = (F_1(u, 0), B(u, 0)v)$.

Note that the equation $(u, v) = H(u, v)$ has been decoupled. That is, one can solve u from the first equation and then substitute the result into the second one.

Proof. By (3.11), $\bar{F}_2(u, 0) = 0$ so that we can write for $t \in [0, 1]$

$$\bar{F}_2(u, tv) = t \int_0^1 \partial_v \bar{F}_2(u, tsv) v ds. \tag{3.16}$$

We consider the following homotopy in $E = U \oplus V$.

$$H(t, u, v) = \left(F_1(u, tv), K(u, tv) \int_0^1 \partial_v \bar{F}_2(u, tsv) v ds \right), \ t \in [0, 1], (u, v) \in E. \tag{3.17}$$

Recall that $K(u,v)$ is a linear map. By (3.16), $H(1,u,v) = F(u,v)$. If there exists $t \in (0,1]$ such that $H(t,u,v)$ has a fixed point $(u,v) \in \partial E$, the boundary of E relative to C, then $v > 0$ since otherwise $u \in Z_1$; but then $(u,0) \notin \partial E$ because of (3.12). Thus, $v > 0$ and (u,tv) is a fixed point of $F(u,v)$ (cf. (3.16)), a contradiction to (3.15). If $H(0,u,v) = (u,v)$ has a solution $(u,v) \in \partial E$ then $u \in Z_1$ and $v > 0$. By (3.17), we have that $v = K(u,0) \circ \partial_v \bar{F}_2(u,0)v = B(u,0)v$ and v is thus a positive eigenvector to $B(u,0)$ with respect to the eigenvalue 1, a contradiction to (B). Hence, the homotopy (3.17) is well defined on E.

Consequently, by the homotopy invariance,

$$\operatorname{ind}(F,E) = \operatorname{ind}(H(1,\bullet),E) = \operatorname{ind}(H(0,\bullet),E). \tag{3.18}$$

This proves the claim since $H(0,\bullet)$ is exactly H. □

In fact, the assumption (3.15) comes from (B) if V is a small neighborhood of 0. Let us denote $C_2(r) = \{v \in C_2 \ : \ \|v\| < r\}$. We then have the next lemma.

LEMMA 3.4. *Assume* (B). *If $V = C_2(r)$ with r sufficiently small, then* (3.15) *is satisfied.*

Proof. Suppose (3.15) is not true for every small $r > 0$. Then there exist sequences of elements $\{u_n\}$,$\{v_n\}$ such that $u_n \in U$, $v_n > 0$, and $\|v_n\| \to 0$ and

$$u_n = F_1(u_n,v_n), \quad v_n = F_2(u_n,v_n) = K(u_n,v_n)\int_0^1 \partial_v \bar{F}_2(u_n,sv_n)v_n ds \quad \text{for all } n.$$

We set $w_n = v_n/\|v_n\|$ and find that

$$w_n = K(u_n,v_n)\int_0^1 \partial_v \bar{F}_2(u_n,sv_n)w_n ds \quad \text{for all } n.$$

By the compactness and continuity of K and $\partial_v \bar{F}_2(\bullet)$, we easily see that $\{w_n\}$ is precompact. Also, the compactness of F_1 implies $\{u_n\}$ is precompact. Without loss of generality, we suppose that $u_n \to u$ and $w_n \to w > 0$. It follows that $u = F_1(u,0)$ and $w = B(u,0)w$. Hence, $u \in Z_1$ and $w > 0$ is an eigenvector of $B(u,0)$ to the eigenvalue 1, a contradiction to (B). The proof is complete. □

We now compute the index of $H(u,v) = (F_1(u,0), B(u,0)v)$. We state the result in a slightly more general form as follows.

PROPOSITION 3.4. *Let $B : U \to L(X_2)$ such that $B(u)$ is a linear operator for every $u \in U$. Assume that $H(u,v) = (F_1(u,0), B(u)v)$ is completely continuous. We then have*

(i) $\operatorname{ind}(H, U \oplus V) = \operatorname{ind}(F_1|_U, U)$ *if for any $u \in Z_1$, $B(u)$ has no positive eigenvector to an eigenvalue greater than one;*

(ii) $\operatorname{ind}(H, U \oplus V) = 0$ *if there exists an element $p \in C_2$ such that, for any $u \in Z_1$, we have*

$$v \neq B(u)v + tp \quad \text{for all } v \in \partial V \quad \text{for all } t > 0. \tag{3.19}$$

Proof. (i). Consider the following homotopy:

$$G(t,u,v) = (F_1(u,0), tB(u)v), \quad t \in [0,1]. \tag{3.20}$$

If $G(t,u,v) = (u,v)$ for some $(u,v) \in \partial E$ and $t \in [0,1]$, then obviously $u \in Z_1$ and $t > 0$ and $v > 0$. But this means that v is a positive eigenvector to the eigenvalue

$t^{-1} > 1$ of $B(u)$ and contradicts our assumption. Thus, by the homotopy invariance of the degree

$$\operatorname{ind}(F, E) = \operatorname{ind}(H(0, \bullet), E) = \operatorname{ind}(G(1, \bullet), E) = \operatorname{ind}(G(0, \bullet), E).$$

However, $G(0, \bullet)$ can be viewed as the product of two maps, $G_1 \equiv F_1|_U$ on U and $G_2 \equiv 0$ on V. Since $\operatorname{ind}(0, V) = +1$, by the product theorem of Leray ([24, Theorem 13.F]) we have

$$\operatorname{ind}(F, E) = \operatorname{ind}(G_1, U) \times \operatorname{ind}(G_2, V) = \operatorname{ind}(F_1|_U, U).$$

Case (ii). We consider the following homotopy:

$$G(t, u, v) = (F_1(u, 0), B(u)v + tp), \quad t \geq 0. \tag{3.21}$$

If $(u, v) \in \partial E$ is a solution of $G(t, u, v) = (u, v)$, then $u \in Z_1$ and $u > 0$. Moreover, $u \in \partial V$ satisfies $v = B(u)v + tp$. This contradicts (B) if $t = 0$ and (3.19) if $t > 0$. Hence, the above homotopy (3.21) is well defined on E.

However, for t large, since $(u, v) \to B(u)v$ is compact, obviously $G(t, u, v)$ has no solution in the bounded set E and therefore $\operatorname{ind}(G(t, \bullet), E)$ must be zero for t large. By homotopy invariance we have that

$$\operatorname{ind}(F, E) = \operatorname{ind}(G(0, \bullet), E) = 0.$$

This completes the proof. □

Since $F_2(u, \bullet)$ maps V into the positive cone C_2 and $F_2(u, 0) = 0$ we easily see that $B(u, 0)$ is a positive linear map. Under a slightly stronger assumption, but often verified in applications, we have the following consequence of Proposition 3.3 and Proposition 3.4

PROPOSITION 3.5. *In addition to* (B), (3.11), (3.12), *and* (3.15), *if we assume further that* $B(u, 0)$ *is strongly positive for all* $u \in Z_1$, *then*

(i) $\operatorname{ind}(F, U \oplus V) = \operatorname{ind}(F_1, U)$ *if for any* $u \in Z_1$, *the spectral radius* $r(B(u, 0)) < 1$.

(ii) $\operatorname{ind}(F, U \oplus V) = 0$ *if for any* $u \in Z_1$, *the spectral radius* $r(B(u, 0)) > 1$.

Proof. We need only to show that the assumptions in (i) and (ii) imply accordingly those of Proposition 3.4 but these facts are just consequences of (ii) and (iv), respectively, of [2, Theorem 3.2]. In particular, by (iv) of [2, Theorem 3.2], we can take any $p > 0$ in C_2 and see that (3.19) is satisfied if $r(B(u, 0)) > 1$. □

Remark 3.5. In fact, the existence of an element p required by (ii) can be guaranteed without the assumption that $B(u, 0)$ is strongly positive. For example, if there exists a demi-interior element p in C_2 (see [4, 3]), then by using the fact that $r(B(u, 0))$ is also an eigenvalue with positive eigenvector for $(B(u, 0))^*$ one can show that (3.19) holds. Another example is the special case when $Z_1 = \{u\}$, i.e., a singleton, then one can take p to be the positive eigenvector to an eigenvalue $\lambda > 1$ of $B(u, 0)$. Indeed, suppose that there is a solution $x > 0$ for some $t > 0$ for $x = B(u, 0)x + tp$. Then there exists a real $\sigma_0 \geq 0$ such that $x \geq \sigma_0 p$ but $x \not\geq \sigma p$ for $\sigma > \sigma_0$. Hence,

$$x = B(u, 0)x + tp \geq B(u, 0)(\sigma_0 p) + tp \geq (\sigma_0 + t)p,$$

which contradicts the maximality of σ_0.

Remark 3.6. If $X_1 = \{0\}$, then we can identify X_2 with $X = \{0\} \oplus X_2$ and the element 0 of X_2 with Z_1. If $K = I$ the identity map, the partial derivative $\partial_v F_2(0, 0)$

can be identified with the right derivative $(F_2)'_+(0)$ in the direction of the cone C_2 (see [2]). In this way, the results of Proposition 3.4 and Proposition 3.5 are generalizations of those in [2, section 13].

We are now ready to give the next proof.

Proof of Theorem 3.1. Let $\beta \in \mathcal{B}$, $\beta \neq \emptyset$. We set $\alpha = M \setminus \beta$. By Lemma 3.4 and (E_1), we can find $r > 0$ sufficiently small such that with $V_\alpha = C_\alpha(r)$ such that $F(u) = u$ has no solution of the form $u = (u_\beta, u_\alpha)$ with $u_\alpha \in V_\alpha$ and $\|u_\alpha\| = r$. We may assume that $U_\beta \subset W_\beta$ and $V_\alpha \subset W_\alpha$.

By Proposition 3.5, $\text{ind}(F, U_\beta \oplus V_\alpha) = \sigma(\beta) i(\beta)$. This gives (i).

Moreover, it is obvious that we can choose U_β, V_α such that the open sets $U_\beta \oplus V_\alpha$ ($\beta \neq M$, $\alpha = M \setminus \beta$) and $U_\emptyset$ are disjoint sets which contain all but positive fixed points of F in Ω. Now, by the additvity property of indices, if (3.8) holds, then the set of positive fixed points of F in Ω must be nonempty. □

3.2. Existence of semitrivial solutions. We see that (3.3) is of the form studied in section 3.1 (see (3.4)). Following the notations in that section, we set $M = \{0, \dots, m\}$ and

$$Z_\beta = \{\vec{u} = (u_0, \dots, u_m) \ : \ F(\vec{u}) = \vec{u}, \quad u_i > 0 \quad \text{if } i \in \beta, \quad u_i = 0 \quad \text{if otherwise}\}.$$

Due to the nonhomogeneity of the boundary condition for u_0 we see that any solution to (2.1) must have $u_0 \neq 0$ so that $Z_\emptyset = \emptyset$. We study first $Z_{\{0\}}$, the set of solutions with $u_0 > 0$ and $u_i = 0$ for all $i \neq 0$. By (F.4), the system reduces to (2.3) so that $Z_{\{0\}}$ consists of exactly the washout solution $\vec{u}_* = (S_*, 0, \dots, 0)$.

In order to apply the results in section 3.1 to compute the face index of $Z_{\{0\}}$ we define

$$K_{\{0\}}(\vec{u}_*) = \text{diag}\{K_1(\vec{u}_*), \dots, K_m(\vec{u}_*)\}, \quad \bar{F}_{\{0\}} = (\bar{F}_1, \dots, \bar{F}_m).$$

We then consider the eigenvalue problem for the operator

$$B_{\{0\}}(\vec{u}_*) = K_{\{0\}}(\vec{u}_*) \circ \partial_{\{0\}} \bar{F}_{\{0\}}. \tag{3.22}$$

By (F.4), $\frac{\partial f_i}{\partial u_j}(x, \vec{u}_*) = 0$ if $j \neq i$. It is easy to see that the eigenvalue problem $B_{\{0\}}(\vec{u}_*)\phi = \lambda\phi$ with $\phi = (U_1, \dots, U_m)$ is equivalent to the system of m equations

$$\begin{cases} -A_i U_i + \text{div}(U_i \Phi_i(S_*) \nabla S_*) = \lambda^{-1} U_i \frac{\partial}{\partial u_i} f_i(x, S_*, 0, \dots, 0), \quad x \in \Omega, \\ \dfrac{\partial U_i}{\partial n} + r_i\left(x, \dfrac{\partial S_*}{\partial n}\right) U_i = 0, \quad x \in \partial\Omega. \end{cases} \tag{3.23}$$

Our principal assumption concerns these eigenvalue problems. In order to guarantee the existence of solutions other than the washout state solution we shall assume that the latter is unstable in its complementary directions. In the light of Theorem 3.1 we shall require that $Z_{\{0\}}$ is "repelling" so that Z_β is nonempty for some $\beta \neq \{0\}$. In terms of the eigenvalue problem (3.23), the condition (E_-) now reads

(E_i) The largest eigenvalue of (3.23) is greater than 1. We say that (E) holds if (E_i) holds for $1 \leq i \leq m$.

Remark 3.7. By using a change of variable as in the proof of Lemma 2.2, we can write (3.23) as

$$\begin{cases} -\text{div}(a_i(x)\nabla U_i) + d_i(x) U_i = \lambda^{-1} g_i(x) U_i \quad \text{on } \Omega, \\ \dfrac{\partial U_i}{\partial n} + R_i(x) U_i = 0 \quad \text{on } \partial\Omega. \end{cases}$$

It is well known that there is only one positive eigenfunction to (3.23) which is the one that corresponds to the largest eigenvalue.

In the same way, Condition (B) of section 3.1 can be restated as follows.

(B′) The largest eigenvalue of (3.23) does not equal 1 for $1 \le i \le m$.

Concerning the eigenvalue problem for $B_{\{0\}}(\vec{u}_*)$ we assert the following (see [11, Lemma 3.2]).

LEMMA 3.8. *If* (B′) *holds, then* 1 *is not an eigenvalue of* $B_{\{0\}}(\vec{u}_*)$ *corresponding to an eigenvector in* X_+. *Moreover, if* (E_i) *holds for some* i *then* $B_{\{0\}}(\vec{u}_*)$ *has an eigenvalue larger than* 1 *with a corresponding eigenvector in* X_+.

Proof. The claims are trivially true if we notice that if ϕ_i, $i > 0$, is the principal eigenfunction of (3.23) to the principal eigenvalue λ_i, then the vector $(u_0, \dots, u_m)$ with $u_j = 0$ if $j \neq i$ and $u_i = \phi_i$ is an eigenvector of $B_{\{0\}}(\vec{u}_*)$ in X_+ corresponding to the eigenvalue λ_i. □

We then have this corollary (see (3.6) for the definition of σ).

COROLLARY 3.6. *Assume* (B′). *If for some* i, $1 \le i \le m$, (E_i) *holds, then*

(i) $\sigma(\{0\}) = 0$;

(ii) *for* $Z_+ = \{u = (u_0, \dots, u_m) \ : \ u = F(u),\ u_i > 0,$ *for some* $i > 0\}$, $\operatorname{ind}(F, Z_+) = 1$;

(iii) *there exists a semitrivial (single-population) equilibrium of* (2.1).

Proof. It is important to point out that, unless $m = 1$, the map $B_{\{0\}}(\vec{u}_*)$ is positive but not strongly positive (otherwise, the positive eigenvector must be unique by the Krein–Rutman theorem). However, since $Z_{\{0\}}$ is a singleton, Lemma 3.8 and Remark 3.2 can be used here to compute $\sigma(\{0\})$. Hence, if (E_i) holds for some i, then (E'_-) is verified so that $\sigma(\{0\}) = 0$.

For item (ii), it is clear that the set of fixed points of F in C_R (see Theorem 2.4) is the union of Z_+ and $Z_{\{0\}}$. By item (i) of Theorem 3.1, Theorem 2.4, and the additivity property of index we have

$$1 = \operatorname{ind}(F, C_R) = i(F, Z_{\{0\}}) + i(F, Z_+) = \sigma(\{0\})i(\{0\}) + i(F, Z_+) = i(F, Z_+).$$

The last equality comes from the fact that $\sigma(\{0\}) = 0$ as we just proved above. Finally, letting $i > 0$, we can easily see that (iii) is a direct consequence of (ii) setting $u_j = 0$ for all $j > 0$ and $j \neq i$. □

3.3. The case of two species. We now turn attention to the two-species case, that is, $m = 2$. It is assumed that for $i = 1, 2$, the principal eigenvalue of the eigenvalue problem (3.23) is greater than 1. Corollary 3.6 then implies the existence of at least one single-population equilibrium for each of the two populations. Thus, $Z_{\{0,1\}}$ and $Z_{\{0,2\}}$, the sets of single-population equilibria for which $u_1 > 0$ or $u_2 > 0$, are nonempty.

Let $\beta = \{0, 1\}$ and $\alpha = \{0, 1, 2\} \backslash \beta = \{2\}$. We define (see (3.1), (3.2))

$$F_\beta(\psi) = (F_0(\psi), K_1(\psi) \circ \bar{F}_1(\psi)),\ F_\alpha(\psi) = K_2(\psi) \circ \bar{F}_2(\psi).$$

Then $F(\psi) = (F_\beta(\psi), F_\alpha(\psi))$. In order to apply the results in section 3.1 to compute the local index of Z_β we need to consider eigenvalue problem for the operator

$$B_\alpha(u) = K_\alpha(u) \circ \partial_\alpha \bar{F}_\alpha(u) = K_2(u) \circ \partial_2 \bar{F}_2(u) \tag{3.24}$$

for $u = (u_0, u_1, 0) \in Z_\beta$. Note that $K_2(u)$ is strongly positive. By (F.3) $B_\alpha(u)$ is also a strongly positive operator. It is easy to see that the eigenvalue problem $B_\beta(u)\phi = \lambda\phi$

is equivalent to the following problem:

$$\begin{cases} -A_2\phi + \operatorname{div}(\phi\Phi_2(u_0)\nabla u_0) = \lambda^{-1}\phi\frac{\partial}{\partial u_2}f_2(x,u_0,u_1,0), & x\in\Omega,\\ \dfrac{\partial\phi}{\partial n} + r_i\left(x,\dfrac{\partial u_0}{\partial n}\right)\phi = 0, & x\in\partial\Omega. \end{cases}$$

Using a similar analysis for the case $\beta=\{0,2\}$, we are led to the following.

Let us denote $Z_i = Z_{\{0,i\}}$, and the elements of Z_i by $\hat{U}_i$, that is, $\hat{U}_1=(\hat{u}_{01},\hat{u}_1,0)$ and $\hat{U}_2=(\hat{u}_{02},0,\hat{u}_2)$. Conditions (E_+) and (E_-) of section 3.1 suggest that we consider the following conditions.

(D_+) For each $i=1,2$, and for any $\hat{U}_i\in Z_i$, and $i,j\in\{1,2\}$ with $j\neq i$, the largest eigenvalue of

$$\begin{cases} -A_i\phi + \operatorname{div}(\phi\Phi_i(\hat{u}_{0j})\nabla(\hat{u}_{0j})) = \lambda^{-1}\phi\frac{\partial}{\partial u_i}f_i(x,\hat{U}_j),\\ \dfrac{\partial\phi}{\partial n} + r_i\left(x,\dfrac{\partial(\hat{u}_{0j})}{\partial n}\right)\phi = 0 \end{cases} \tag{3.25}$$

is greater than 1.

(D_-) For each i, the eigenvalues of (3.25) are all less than 1.

In biological terms, (D_+) says that every u_1-single population steady state is stable to invasion by u_2 and conversely. (D_-) says that every u_1-single population steady state is unstable to invasion by u_2 and conversely. In either case, we might expect the existence of a positive steady state (u_0,u_1,u_2), that is, a steady state with $u_i>0$ for $i=1,2$. The main result of this section asserts that this is the case.

THEOREM 3.7. *Let $m=2$ and assume (E_i) holds for $i=1,2$ and either (D_+) or (D_-) holds. Then the system (2.1) has at least one positive solution.*

Proof. By Corollary 3.6, $\sigma(\{0\})=0$. Let $\beta=\{0,1\}$ or $\{0,2\}$. If (D_+) (resp., (D_-)) holds, then (E_+) (resp., (E_-)) of section 3.1 holds so that $\sigma(\beta)=1$ (resp., 0). Note also that, in this case, $B_\alpha(\hat{U})$ is strongly positive for any $\hat{U}\in Z_i$.

Since (E_i) holds for $i=1,2$, by (ii) of Corollary 3.6 we see that the face index $i(\beta)=1$. In fact, $i(\beta)$ is just the local index of Z_β as the set of positive fixed points of F restricted to the face X_β so that we can apply (ii) of Corollary 3.6.

Therefore, the right-hand side of (3.8) can be computed as follows.

$$\sum_{\beta=\{0\},\{0,1\},\{0,2\}}\sigma(\beta)i(\beta) = \begin{cases} 2 & \text{if } (D_+) \text{ holds},\\ 0 & \text{if } (D_-) \text{ holds}. \end{cases}$$

In both cases this quantity is not equal to $\operatorname{ind}(F,C_R)=1$. By (ii) of Theorem 3.1 our theorem now follows. ∎

Acknowledgments. The author would like to thank the anonymous referees for valuable constructive comments which led to many improvements of this work.

REFERENCES

[1] W. ALLEGRETTO, H. XIE, AND S. YANG, *Properties of solutions for a chemotaxis system*, J. Math. Biol., 35 (1997), pp. 949–966.

[2] H. AMANN, *Fixed point equations and nonlinear eigenvalue problems in ordered Banach spaces*, SIAM Rev., 18 (1976), pp. 620–709.

[3] E. N. DANCER AND Y. DU, *Positive solutions for a three-species competition system with diffusion–Part* I *and Part* II, Nonlinear Anal., 24 (1995), pp. 337–373.

[4] E. N. Dancer, *On positive solutions of some pairs of differential equations*, Trans. Amer. Math. Soc., 284 (1984), pp. 729–743.
[5] L. Dung, *Coexistence with Chemotaxis. Part* II*: Periodic Solutions*, CDSNS report, Center for Dynamical Systems and Nonlinear Studies, Atlanta, GA, 1998.
[6] L. Dung, *Global attractors and steady state solutions for a class of reaction diffusion systems*, J. Differential Equations, 147 (1998), pp. 1–29.
[7] L. Dung, *Ultimately uniform boundedness of solutions and gradients for degenerate parabolic systems*, Nonlinear Anal., 39 (2000), pp. 157–171.
[8] L. Dung, *Dissipativity and global attractors for a class of quasilinear parabolic systems*, Comm. Partial Differential Equations, 22 (1997), pp. 413–433.
[9] L. Dung, *On steady state solutions for a class of reaction diffusion systems*, Canadian Appl. Math. Quart., 6 (1997), pp. 1–18.
[10] L. Dung and H. Smith, *Steady states of models of microbial growth and competition with chemotaxis*, J. Math. Anal. Appl., 229 (1999), pp. 295–318.
[11] L. Dung and H. L. Smith, *On a parabolic system modelling microbial competition in an unmixed bio-reactor*, J. Differential Equations, 130 (1996), pp. 59–91.
[12] A. Friedman, *Partial Differential Equations*, Holt, Rinehart and Winston, New York, 1969.
[13] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, 1983.
[14] K. P. Hadeler, F. Rothe, and H. Vogt, *Stationary solutions of reaction-diffusion equations*, Math. Methods Appl. Sci., 1 (1979), pp. 418–431.
[15] M. Herrero and J. Velázquez, *Chemotactic collapse for the Keller-Segel model*, J. Math. Biol., 35 (1996), pp. 177–194.
[16] W. Jäger and S. Luckhaus, *On explosion of solutions to a system of partial differential equations modelling chemotaxis*, Trans. Amer. Math. Soc., 329 (1992), pp. 819–824.
[17] E. F. Keller and L. A. Segel, *Initiation of slime mold aggregation viewed as an instability*, J. Theory Biol., 26 (1970), pp. 399–415.
[18] F. Kelly, K. Dapsis, and D. Lauffenburger, *Effect of bacterial chemotaxis on dynamics of microbial competition*, Microbial Ecology, 16 (1988), pp. 115–131.
[19] D. Lauffenburger, *Quantitative studies of bacterial chemotaxis and microbial population dynamics*, Microbial Ecology, 22 (1991), pp. 175–185.
[20] D. Lauffenburger, R. Aris, and K. Keller, *Effects of cell motility and chemotaxis on microbial population growth*, Biophys. J., 40 (1982), pp. 209–219.
[21] C. S. Lin, W. M. Ni, and I. Takagi, *Large amplitude stationary solutions to a chemotaxis system*, J. Differential Equations, 72 (1988), pp. 1–27.
[22] R. Schaff, *Stationary solutions of chemotaxis systems*, Trans. Amer. Math. Soc., 292 (1985), pp. 531–556.
[23] X. Wang, *Qualitative Behavior of Solutions of a Chemotactic Diffusion System: Effects of Motility and Chemotaxis and Dynamics*, preprint, 1998.
[24] E. Zeidler, *Nonlinear Functional Analysis and Its Applications*, Springer-Verlag, New York, 1993.

SIAM J. MATH. ANAL.
Vol. 32, No. 3, pp. 522–537

HARMONIC MOMENTS AND AN INVERSE PROBLEM FOR THE HEAT EQUATION*

MISHIO KAWASHITA†, YAROSLAV V. KURYLEV‡, AND HIDEO SOGA†

Abstract. In the paper, we study an inverse problem for the heat equation. We introduce a class of bilinear forms on the space of harmonic polynomials (called harmonic moments), which are represented by the Dirichlet-to-Neumann map. We investigate the uniqueness, stability, and reconstruction of the inverse problem.

The inverse data are given in the terms of the bilinear forms and can be exchanged into the data of the Dirichlet-to-Neumann map. The reconstruction (of the density) is accomplished in two different ways: one is due to the idea of the mollifier and the other to the representation by the Carleman kernel in the complex analysis. The error terms are estimated depending on the degree of the harmonic polynomials. We estimates norms of the data on an arbitrary time interval by the norms on some fixed interval (e.g., $0 < t < 2$).

Key words. inverse problem, heat equation, harmonic moments, reconstruction of a coefficient

AMS subject classifications. 35R30, 35K05, 80A23, 41A27

PII. S0036141099353035

1. Introduction.

This paper is devoted to the solution of the inverse boundary problem for the heat equation. Let Ω be a connected bounded domain in $\mathbb{R}^n$ $(n \geq 2)$ with C^l ($l = 2, 3, \dots$) boundary Γ. Consider the mixed problem for the heat equation

$$\begin{cases} (\rho(x)\partial_t - \triangle)u(t,x) = 0 & \text{in} \quad (0,+\infty)\times\Omega, \\ u(t,x') = f(t,x') & \text{on} \quad (0,+\infty)\times\Gamma, \\ u(0,x) = 0 & \text{on} \quad \Omega. \end{cases} \tag{1.1}$$

The density $\rho(x)$ is a $C^{l+\sigma}$ $(0 < \sigma < 1)$ function on $\overline{\Omega}$ satisfying

$$0 < \rho_1 \leq \rho(x) \leq \rho_2 (< +\infty). \tag{1.2}$$

The inverse data used in the paper is a set of normal derivatives $\frac{\partial u^p}{\partial \nu}|_{(0,2)\times\Gamma}$, where u^p is the solution of (1.1) with

$$f(t,x') = \chi(t)\, p(x'). \tag{1.3}$$

Here $\chi(t)$ is a (arbitrary) fixed C^∞ function satisfying $0 \leq \chi(t) \leq 1$ in $\mathbb{R}$, $\chi(t) = 1$ for $t \geq 1$ and $\chi(t) = 0$ for $t \leq 1/2$. The function $p(x')$ in (1.3) is the boundary value of a harmonic polynomial $p(x)$ (i.e., $\triangle p = 0$).

We assume that $\frac{\partial u^p}{\partial \nu}|_{(0,2)\times\Gamma}$ or, more precisely,

$$\int_\Gamma \int_0^t \frac{\partial u^p}{\partial \nu}(s,x')\, q(x')\, dx' ds, \quad 0 < t < 2, \tag{1.4}$$

*Received by the editors February 22, 1999; accepted for publication (in revised form) January 21, 2000; published electronically September 15, 2000. The second and third authors were partly supported by EPSRS grant GR/L66472.

http://www.siam.org/journals/sima/32-3/35303.html

†Faculty of Education, Ibaraki University, Mito Ibaraki, 310-8512, Japan (kawasita@mito.ipc.ibaraki.ac.jp, soga@mito.ipc.ibaraki.ac.jp).

‡Department of Mathematical Science, Loughborough University, Loughborough, Leics, LE11 3TU, UK (Y.V.kurylev@lboro.ac.uk).

are given for all sources f of form (1.3) with $p \in HP^m$, where

$$HP^m = \{\text{harmonic polynomials of degree} \le m\} \qquad (m = 0, 1, 2, \dots),$$

and all $q \in HP^m$.

In this paper we describe algorithms for an approximate reconstruction of ρ given approximate integrals (1.4) with $p, q \in HP^m$, $m = 0, 1, 2, \dots$. The algorithms lead to explicit formulae for an approximate solution of the inverse problem under consideration together with an error estimate in the corresponding reconstruction procedure. These error estimates depend upon the parameter m and a "measurement" error, i.e., an error in the inverse data (1.4) and of the logarithmic character. The algorithms are described in section 3, where we do not use a quasi-analytic continuation of the inverse data, and in section 4, where we utilize results on quasi-analytic continuation (see, e.g., [Car], [L]). The corresponding formulae together with the error estimates are given in Theorems 3.4 and 4.4.

The method used in the paper is the parabolic analog of the moments method introduced in [K-S] for the solution of the inverse boundary spectral problem for the elliptic operator $A_\rho = -\rho^{-1}\triangle$. Its main idea is to utilize the generating properties of the products of harmonic polynomials. These polynomials belong to the null-space of the operator A_ρ. The considered generating properties are, in fact, an algebraic version of the well-known fact (see, e.g., [Cal]) that the linear combinations of the products of harmonic functions are dense in $L^2(\Omega)$.

This fact was extensively used for solving inverse boundary problems (see [S-U] for the pioneering work in this direction). In the parabolic case the study of the inverse problems for the system (1.1) and even for some more general parabolic equations was carried out by a number of authors. The main results dealt with the question of uniqueness and stability in the identification of the unknown coefficient(s) via various sets of the inverse data on the boundary. A very good introduction to this area together with a number of advanced results is given in [Is, Chapter 9]. However, we would like to stress that our main goal is not only to obtain stability estimates but also to develop some reconstruction procedures and to obtain stability estimates for these procedures. In its turn such procedures may show useful for the numerical solution of inverse boundary problems. Indeed, the acoustic variant of the moments method was successfully used in [K-P] for the numerical solution of some model inverse problems.

In the note [K-K-S] we have partially announced the results (only in sections 2 and 3) with an explanation of the key idea of the proofs.

In the present paper we confine ourselves to the problems of stability in Hölder classes. However, the method may also be used (see, e.g., [K-P], [K-S]) to analyze stability in Sobolev classes. The case of the other types of boundary conditions may be considered as well. This will be done elsewhere.

2. Direct problem and uniqueness of inverse problem. Consider the mixed problem (1.1) and denote by $u^p(t, x)$ its solution with the source f of form (1.3). When p, $q \in HP^\infty (= \cup_{m=0}^\infty HP^m)$, we define $\Phi_\rho(t; p, q)$ by

$$\Phi_\rho(t; p, q) = \int_\Omega \rho(x) u^p(t, x) \overline{q(x)}\, dx. \tag{2.1}$$

We define also the harmonic moments $M_\rho(p, q)$ corresponding to ρ:

$$M_\rho(p, q) = \int_\Omega \rho(x) p(x) \overline{q(x)}\, dx. \tag{2.2}$$

In this section we discuss some properties of $\Phi_\rho(t;p,q)$ and relations between $\Phi_\rho(t;p,q)$, $M_\rho(p,q)$, and the response operator

$$R_\rho : p(x') \mapsto \frac{\partial u^p}{\partial \nu}|_{(0,+\infty)\times\Gamma}.$$

THEOREM 2.1. *$\Phi_\rho(t;p,q)$ is a bilinear mapping from $HP^\infty \times HP^\infty \to C^0(\overline{\mathbb{R}_+})$ with the following properties:*

(i) $$\left\|\Phi_\rho(\cdot;p,q)\right\|_{C^0(\overline{\mathbb{R}_+})} \le C \left\|p\right\|_{L^2(\Omega)} \left\|q\right\|_{L^2(\Omega)},$$

(ii) $$\left\|\Phi_\rho(\cdot;p,q) - \Phi_{\tilde\rho}(\cdot;p,q)\right\|_{C^0(\overline{\mathbb{R}_+})} \le C \left\|\rho-\tilde\rho\right\|_{C^0(\overline{\Omega})} \left\|p\right\|_{L^2(\Omega)} \left\|q\right\|_{L^2(\Omega)},$$

(iii) $$\left|\Phi_\rho(t;p,q) - M_\rho(p,q)\right| \le Ce^{-C'\lambda_0 t} \left\|p\right\|_{L^2(\Omega)} \left\|q\right\|_{L^2(\Omega)},$$

where λ_0 is the first eigenvalue of the Dirichlet problem for the Euclidean Laplacian $-\triangle$ on Ω, and where C, C' are positive constants determined from ρ_1, ρ_2; C' does not depend on λ_0.

Note (1). We denote by C, C' different constants which depend upon Ω, ρ_1, and ρ_2. In the case of their dependence upon some other parameters we note this dependence explicitly.

Note (2). Hereafter we denote the norm $\left\|\cdot\right\|_{L^2(\Omega)}$ only by $\left\|\cdot\right\|$.

Proof. Consider the mixed problem

$$\begin{cases} \partial_t w(t,x) - \rho^{-1}\triangle w(t,x) = 0 & \text{in} \quad (0,+\infty)\times\Omega, \\ w(t,x') = 0 & \text{on} \quad (0,+\infty)\times\Gamma, \\ w(0,x) = f(x) & \text{on} \quad \Omega, \end{cases}$$

and denote by $\{\lambda_j\}_{j=1,2,\dots}$, $\lambda_1 < \lambda_2 \le \cdots$ the eigenvalues of the self-adjoint differential operator $-\rho^{-1}\triangle$ with the domain $H^2(\Omega)\cap H_0^1(\Omega)$ in the Hilbert space $L^2_\rho(\Omega)$ with the inner product $(f,g)_\rho \equiv \int_\Omega \rho(x) f(x)\overline{g(x)}\,dx$. We can express the solution $w(t,\cdot) = E(t)f$ in the form

$$E(t)f = \sum_{j=1}^{\infty} e^{-\lambda_j t}(f,\varphi_j)_\rho\,\varphi_j, \tag{2.3}$$

where the functions $\{\varphi_j\}_{j=1,2,\dots}$ are the L^2_ρ-orthonormal eigenfunctions corresponding to the eigenvalues λ_j. Note that the first eigenvalue λ_1 is estimated by λ_0, i.e., $\lambda_0\rho_2^{-1} \le \lambda_1 \le \lambda_0\rho_1^{-1}$.

Then we have

$$u^p(t,x) - \chi(t)p(x) = \int_0^t E(t-s)(-\chi'(s)p(x))\,ds. \tag{2.4}$$

Since $L^2_\rho(\Omega) = L^2(\Omega)$ as a set and $\rho_1\left\|f\right\|^2 \le \left\|f\right\|_\rho^2 (= (f,f)_\rho) \le \rho_2\left\|f\right\|^2$, (2.3) yields that

$$\left\|\int_0^t E(t-s)(-\chi'(s)p(x))\,ds\right\| \le (\rho_2/\rho_1)^{1/2} e^{\lambda_1(1-t)} \sup_{s\in\mathbb{R}}\left|\chi'(s)\right| \left\|p\right\|.$$

Thus, we obtain

$$\begin{aligned}
&\left|\Phi_\rho(t;p,q) - \int_\Omega \rho(x)p(x)\overline{q(x)}\,dx\right| \\
&\quad\le \left|(1-\chi(t))\int_\Omega \rho p\overline{q}\,dx\right| + \left|\int_\Omega \rho(u-\chi p)\overline{q}\,dx\right| \\
&\quad\le \left(e^{\lambda_1}\rho_2 + (\rho_2/\rho_1)^{1/2}e^{\lambda_1}\left\|\chi'\right\|_{C^0(\overline{\mathbb{R}_+})}\right)e^{-\lambda_1 t}\left\|p\right\|\left\|q\right\|,
\end{aligned}$$

which proves (i) and (iii).

Denote by $\tilde{u}^p(t,x)$ the solution of (1.1) with $\tilde{\rho}(x)$ instead of $\rho(x)$, and by $\widetilde{E}(t)$ the operator of form (2.3) corresponding to $\tilde{\rho}$. We see that

$$u^p(t,x) - \tilde{u}^p(t,x) = \int_0^t E(t-s)\{(\tilde{\rho}-\rho)\rho^{-1}\partial_s\tilde{u}^p(s,\cdot)\}\,ds.$$

Therefore (1.2) implies that

$$\left\|u^p(t,x) - \tilde{u}^p(t,x)\right\| \le \rho_1^{-1}\,(\rho_2/\rho_1)^{1/2}\,\left\|\rho-\tilde{\rho}\right\|_{C^0(\overline{\Omega})}\int_0^t \left\|\partial_s\tilde{u}(s,\cdot)\right\|\,ds. \tag{2.5}$$

From (2.4) it follows that $\partial_s\tilde{u}^p(s,\cdot) = -\int_0^s \partial_t\widetilde{E}(s-t)[\chi'(t)p]\,dt = \chi'(s)p - \int_0^s \widetilde{E}(s-t)\chi''(t)p\,dt$, which yields that

$$\begin{aligned}
\left\|\partial_s\tilde{u}^p(s,\cdot)\right\| &\le \left|\chi'(s)\right|\left\|p\right\| \\
&\quad + (\rho_2/\rho_1)^{1/2}\int_0^1 e^{-\tilde{\lambda}_1(s-t)}\,dt\left\|\chi''\right\|_{C^0(\overline{\mathbb{R}_+})}\left\|p\right\|.
\end{aligned}$$

In the above, $\tilde{\lambda}_1$ is the first eigenvalue of the self-adjoint realization of $-\tilde{\rho}^{-1}\triangle$ on $L^2_{\tilde{\rho}}(\Omega)$. Combining this with (2.5), we obtain

$$\begin{aligned}
\left|\Phi_\rho(t;p,q) - \Phi_{\tilde{\rho}}(t;p,q)\right| &\le \left|\int_\Omega(\tilde{\rho}-\rho)u(t,\cdot)\overline{q}\,dx\right| \\
&\quad + \left|\int_\Omega \tilde{\rho}\{u(t,\cdot)-\tilde{u}(t,\cdot)\}\overline{q}\,dx\right| \\
&\le C\left\|\rho-\tilde{\rho}\right\|_{C^0(\overline{\Omega})}\left\|p\right\|\left\|q\right\|.
\end{aligned}$$

This completes the proof of Theorem 2.1. □

The following lemma shows the relationship between the form $\Phi_\rho(t;p,q)$ and the response operator R_ρ:

LEMMA 2.2. *For any p, $q \in HP^\infty$, we have*

$$\begin{aligned}
\Phi_\rho(t;p,q) &= \int_\Gamma\int_0^t R_\rho[p|_\Gamma](s,x')\,ds\,\overline{q(x')}\,dx' \\
&\quad - \int_0^t \chi(s)\,ds\int_\Gamma p(x')\frac{\partial\overline{q(x')}}{\partial\nu}\,dx'.
\end{aligned}$$

Proof. By integration by parts we have

$$\begin{aligned}\int_\Omega \int_0^t \{(\partial_s - \rho^{-1}\triangle)u^p(s,x)\}\overline{q(x)}\rho(x)\,ds\,dx \\ = \int_\Omega u^p(t,x)\overline{q(x)}\rho(x)\,dx - \int_\Omega u^p(0,x)\overline{q(x)}\rho(x)\,dx \\ - \int_\Omega \int_0^t u^p(s,x)\overline{\triangle q(x)}\,ds\,dx \\ - \int_\Gamma \int_0^t \frac{\partial u^p}{\partial \nu}(s,x')\overline{q(x')}\,ds\,dx' + \int_\Gamma \int_0^t u^p(s,x')\overline{\frac{\partial q}{\partial \nu}(x')}\,ds\,dx'.\end{aligned}$$

As u^p is a solution of (1.1) and $q \in HP^\infty$, the above inequality implies that

$$\begin{aligned}0 = \Phi_\rho(t;p,q) - \int_\Gamma \int_0^t R_\rho[p|_\Gamma](s,x')\overline{q(x')}\,ds\,dx' \\ + \int_\Gamma \int_0^t \chi(s)\,ds\,p(x')\overline{\frac{\partial q}{\partial \nu}(x')}\,dx'.\end{aligned}$$

This proves Lemma 2.2. □

In the inverse problems, generally, we expect to recover ρ by the measurements expressed in terms of the response operator. Lemma 2.2 implies that our original setting of the inverse problem may be reduced to the inversion of the mapping : $\rho \mapsto \Phi_\rho$.

We start with verifying the uniqueness of the inverse problem.

THEOREM 2.3. *If for any p, $q \in HP^\infty$, $\Phi_\rho(t;p,q)$ is equal to $\Phi_{\tilde\rho}(t;p,q)$ on an interval $(1 <) a < t < b$, then ρ coincides with $\tilde\rho$.*

Remark. Lemma 2.2 means that $\Phi_\rho = \Phi_{\tilde\rho}$ if $R_\rho = R_{\tilde\rho}$, and therefore the unique determination of ρ by R_ρ is derived from Theorem 2.3.

The proof of Theorem 2.3 is based on the following lemma.

LEMMA 2.4. *Any polynomial can be expressed as a linear combination of products of the harmonic polynomials.*

For the proof of Lemma 2.4, see [K-S, Proposition 3].

Proof of Theorem 2.3. The solution $u^p(t,x)$ in (1.1) becomes analytic in t for $t > 1$ (cf. (2.3)). Therefore, if $\Phi_\rho(t;p,q) = \Phi_{\tilde\rho}(t;p,q)$ on (a,b), $\Phi_\rho(t;p,q)$ is equal to $\Phi_{\tilde\rho}(t;p,q)$ on $(1,+\infty)$. Hence, by (iii) of Theorem 2.1, we have

$$M_\rho(p,q) = \lim_{t\to\infty} \Phi_\rho(t;p,q) = \lim_{t\to\infty} \Phi_{\tilde\rho}(t;p,q) = M_{\tilde\rho}(p,q), \quad p,q \in HP^\infty.$$

By Lemma 2.4 this implies that

$$\int_\Omega \rho(x)x^\alpha\,dx = \int_\Omega \tilde\rho(x)x^\alpha\,dx$$

for any multi-index α. As the domain Ω is bounded, this implies that $\rho = \tilde\rho$. The proof is complete. □

3. Reconstruction of $\rho(x)$. In this section, we reconstruct $\rho(x)$ approximately, by employing the harmonic moments $M_\rho(p,q)$ with $p,q \in HP^m$, where m is a sufficiently large positive integer.

The reconstruction is based on the fact that the Gaussian distribution

$$(\sqrt{\pi})^{-n}\mu^n \exp(-\mu^2 \left|x\right|^2) = (\sqrt{\pi})^{-n}\mu^n \sum_{k=0}^{\infty} \frac{1}{k!}(-\mu^2 \left|x\right|^2)^k$$

tends to the Dirac δ-function as $\mu \to +\infty$. (Note that $\int_{\mathbb{R}^n} \mu^n \exp(-\mu^2|y|^2)\, dy = (\sqrt{\pi})^n$ for any $\mu > 0$.) Namely, we use the following.

LEMMA 3.1. *Let*

$$\delta_\mu^m(x) = (\sqrt{\pi})^{-n}\mu^n \sum_{k=0}^{m/2} \frac{(-\mu^2 \left|x\right|^2)^k}{k!},$$

where $\mu \geq 1$ and m is a positive even integer. Then for any $\rho(x) \in C^{l+\sigma}(\overline{\Omega})$ $(0 \leq l \leq m/2,$ integer, $0 < \sigma < 1)$ we have

$$\left\| \rho(\cdot) - \int_\Omega \delta_\mu^m(\cdot - y)\rho(y)\, dy \right\|_{C^l(\mu^{-1/2})} \leq C \left\|\rho\right\|_{C^{l+\sigma}} \{\mu^{-\sigma} + (C'\mu)^{m+n+2} m^{-m/2+l}\},$$

where C, C' are independent of ρ, μ, and m. Here we denote by $\left\|\rho\right\|_{C^{l+\sigma}}$ the $C^{l+\sigma}$-norm of ρ in $\overline{\Omega}$ and by $\left\|\rho\right\|_{C^l(\epsilon)}$ the C^l-norm of ρ in $\overline{\Omega_\epsilon}$. In its turn, $\Omega_\epsilon = \{x \in \Omega | \text{dist } (x, \Gamma) > \epsilon\}$.

Proof. For $f \in C^l(\overline{\Omega})$ we denote by Df the C^l-continuation of f onto $\mathbb{R}^n$ with supp $[Df] \subset \{x | \text{dist } (x, \Omega) < 1\}$ and $\left\|Df\right\|_{C^l(\mathbb{R}^n)} \leq 2\left\|f\right\|_{C^l}$. As $\int_{\mathbb{R}^n} \delta_\mu^\infty\ (x)\, dx = 1$, then for any $\left|\alpha\right| \leq l$

$$\partial_x^\alpha \rho(x) - \partial_x^\alpha \int_\Omega \delta_\mu^m(x-y)\rho(y)\, dy = \int_{\mathbb{R}^n} \delta_\mu^\infty(y)(\partial_x^\alpha \rho(x) - \partial_x^\alpha D\rho(x-y))\, dy$$
$$+ \int_{\mathbb{R}^n \setminus \Omega} \partial_x^\alpha \delta_\mu^\infty(x-y) D\rho(y)\, dy + \int_\Omega (\partial_x^\alpha \delta_\mu^\infty(x-y) - \partial_x^\alpha \delta_\mu^m(x-y))\rho(y)\, dy$$
$$= I_1 + I_2 + I_3.$$

Note that $\delta_\mu^\infty(x) - \delta_\mu^m(x) = (\sqrt{\pi})^{-n}\mu^n(-1)^{m/2+1} F_m(\mu^2|x|^2)$, where $F_m(X) = X^{m/2+1} R_m(X)$ and

$$R_m(X) = \frac{1}{(m/2)!} \int_0^1 (1-\theta)^{m/2} e^{-\theta X}\, d\theta.$$

Then, $\sup_{0 \leq \theta \leq 1} \theta^{q_2} e^{-\theta X} \leq q_2! X^{-q_2}$ for any nonnegative integer q_2, and we have

$$|\partial_X^{q_1+q_2} R_m(X)| \leq \frac{q_2! X^{-q_2}}{(m/2)!} \int_0^1 (1-\theta)^{m/2} \theta^{q_1}\, d\theta = \frac{q_1! q_2!}{(m/2+q_1+1)!} X^{-q_2},$$

which yields

$$|\partial_X^q F_m(X)| \leq \frac{2^q}{(m/2+1-q)!} X^{m/2+1-q} \quad \text{for any } X \geq 0, q = 0, 1, \ldots, m/2+1.$$

Using this estimate, we can get

$$|\partial_x^\alpha((\partial_X^q F_m)(\mu^2|x|^2))| \le \frac{2^q 4^{|\alpha|}|\alpha|!\mu^{|\alpha|}}{(m/2+1-q-|\alpha|)!}(\mu|x|)^{2(m/2+1-q)-|\alpha|}$$
$$\text{for any } x \in \mathbb{R}^n, q+|\alpha| \le m/2+1, \mu > 0$$

by induction in $|\alpha|$. Hence, for $|\alpha| \le m/2+1$ we obtain

$$\left|\partial_x^\alpha(\delta_\mu^\infty(x) - \delta_\mu^m(x))\right| \le \frac{4^{|\alpha|}|\alpha|!}{(m/2+1-|\alpha|)!}(\sqrt{\pi})^{-n}\mu^{n+|\alpha|}(\mu\left|x\right|)^{m+2-|\alpha|}.$$

This implies that

$$\begin{aligned}\left|I_3\right| &\le (\sqrt{\pi})^{-n}\mu^{n+|\alpha|}\frac{4^{|\alpha|}\,|\alpha|!}{(m/2+1-|\alpha|)!}(\mu r_0)^{m+2-|\alpha|}\int_\Omega \rho(y)\,dy \\ &\le (\sqrt{\pi})^{-n}4^{|\alpha|}\,|\alpha|!\mu^n\frac{(m/2)^{|\alpha|}}{r_0^{|\alpha|}}\frac{(\mu r_0)^{(m+2)}}{(m/2)!}\int_\Omega \rho(y)\,dy \\ &\le Cr_0^2\,|\alpha|!\mu^{n+m+2}\left(\frac{2m}{r_0}\right)^{|\alpha|}\frac{(2er_0^2)^{m/2}}{m^{(m+1)/2}}\int_\Omega \rho(y)\,dy,\end{aligned}$$

where $r_0 = \operatorname{diam}(\Omega)$ and we use Stirling formula to estimate $(\frac{m}{2})!$ for sufficiently large m.

Since $|x-y| \ge \mu^{-1/2}$ holds if $x \in \Omega_{\mu^{-1/2}}$ and $y \in \mathbb{R}^n \setminus \Omega$, we obtain

$$\begin{aligned}\left|I_2\right| &\le C\left\|\rho\right\|_{C^0}\int_{\mathbb{R}^n\setminus\Omega}\mu^{n+|\alpha|}(1+\mu\left|x-y\right|)^{|\alpha|}e^{-\mu^2|x-y|^2}\,dy \\ &\le C\left\|\rho\right\|_{C^0}\mu^{|\alpha|}e^{-\mu/2}\int_{\mathbb{R}^n}(1+|y|)^{|\alpha|}e^{-|y|^2}\,dy.\end{aligned}$$

At last

$$\begin{aligned}\left|I_1\right| &\le \int_{|y|\le\mu^{-1/2}}\delta_\mu^\infty(y)\left\|\rho\right\|_{C^{l+\sigma}}|y|^\sigma\,dy \\ &\qquad + \int_{|y|>\mu^{-1/2}}\delta_\mu^\infty(y)\{\left\|D\rho\right\|_{C^l(\mathbb{R}^n)} + \left\|\rho\right\|_{C^l(\mathbb{R}^n)}\}\,dy \\ &\le \mu^{-\sigma}\int_{\mathbb{R}^n}|y|^\sigma e^{-|y|^2}\,dy\left\|\rho\right\|_{C^{l+\sigma}} + Ce^{-\mu/2}\int_{\mathbb{R}^n}e^{-|y|^2/2}\,dy\left\|\rho\right\|_{C^l}.\end{aligned}$$

Combining the above estimates, we obtain Lemma 3.1.

The proof is complete. □

We fix a basis $\{p_i^m\}_{i=1,\dots,N(m)}$ in HP^m so that $\{p_i^m\} \subset \{p_i^{m+1}\}$. Lemma 2.4 means that any x^α $(\left|\alpha\right| \le m)$ is expressed in the form

$$x^\alpha = \sum_{i,j=1}^{N(m)} C_{i,j}^\alpha p_i^m(x)\overline{p_j^m(x)}.$$

The function $\delta_\mu^m(x-y)$ in Lemma 3.1 is then decomposed into a sum of polynomials $x^\alpha y^\beta$:

$$\delta_\mu^m(x-y) = \sum_{|\alpha+\beta|\le m} C_{m,\alpha,\beta}\,\mu^{n+|\alpha+\beta|}x^\alpha y^\beta.$$

Therefore the integral $\int_\Omega \delta_\mu^m(x-y)\rho(y)\,dy$ is then represented in terms of the harmonic moments $M_\rho(p,q)$:

$$\int_\Omega \delta_\mu^m(x-y)\rho(y)\,dy = \sum_{|\alpha+\beta|\le m} C_{m,\alpha,\beta}\,\mu^{n+|\alpha+\beta|}x^\alpha \sum_{i,j=1}^{N(m)} C_{ij}^\beta M_\rho(p_i^m,p_j^m).$$

This representation together with Lemma 3.1 implies the possibility of an approximate reconstruction of ρ in terms of the harmonic moments. For this end we introduce the polynomials $Q_\mu^m(x;M)$, where M is a bilinear form on $HP^\infty \times HP^\infty$:

$$Q_\mu^m(x;M) = \sum_{|\alpha+\beta|\le m} C_{m,\alpha,\beta}\mu^{n+|\alpha+\beta|}x^\alpha \sum_{i,j=1}^{N(m)} C_{ij}^\beta M(p_i^m,p_j^m). \tag{3.1}$$

THEOREM 3.2. (i) *The mapping : $M \mapsto Q_\mu^m(x;M)$ is continuous in the following sense:*

$$\begin{aligned}&\left\|Q_\mu^m(\cdot;M) - Q_\mu^m(\cdot;\widetilde{M})\right\|_{C^l}\\ &\le C\mu^n e^{C'\mu^2}\left\|M-\widetilde{M}\right\|_m \max_{|\beta|\le m}\sum_{i,j=1}^{N(m)}\left|C_{ij}^\beta\right|\left\|p_i^m\right\|\left\|p_j^m\right\|,\end{aligned} \tag{3.2}$$

where

$$\left\|M-\widetilde{M}\right\|_m = \sup\left\{\frac{\left|M(p,q)-\widetilde{M}(p,q)\right|}{\|p\|\,\|q\|};\quad p,q\in HP^m\right\}.$$

(ii) *Let $\rho(x)\in C^{l+\sigma}(\overline{\Omega})$ $(0\le l\le m/2,$ integer $0<\sigma<1)$. Then we have*

$$\begin{aligned}&\left\|\rho(\cdot)-Q_\mu^m(\cdot,M_\rho)\right\|_{C^l(\mu^{-1/2})}\\ &\qquad\le C\left\|\rho\right\|_{C^{l+\sigma}}\{\mu^{-\sigma}+(C'\mu)^{(m+n+2)}m^{-m/2+l}\}.\end{aligned} \tag{3.3}$$

Here, the constants C, C' are independent of ρ, μ, and m.

Proof. The estimate (3.3) of Theorem 3.2 follows from Lemma 3.1 immediately. The estimate (3.2) is also easily checked:

$$\begin{aligned}\left\|Q_\mu^m(\cdot;M)-Q_\mu^m(\cdot;\widetilde{M})\right\|_{C^l} \le &\sum_{|\alpha+\beta|\le m}\left|C_{m,\alpha,\beta}\right|\mu^{n+|\alpha+\beta|}r_1^{|\alpha|}\\ &\left\|M-\widetilde{M}\right\|_m\sum_{i,j=1}^{N(m)}\left|C_{ij}^\beta\right|\left\|p_i^m\right\|\left\|p_j^m\right\|,\end{aligned}$$

where $r_1=\max\{|x|\,;\,x\in\Omega\}$. Since each $C_{m,\alpha,\beta}$ is the coefficient of the expansion of $\delta_\mu^m(x-y)$, we obtain

$$\sum_{|\alpha+\beta|\le m}\left|C_{m,\alpha,\beta}\right|\mu^{n+|\alpha+\beta|}r_1^{|\alpha|} \le (\sqrt{\pi})^{-n}\,\mu^n e^{n\mu^2(r_1+1)^2},$$

which implies the estimate (3.2). ∎

From Theorem 3.2 and Theorem 2.1, we have the following.

COROLLARY 3.3. *Let $\Phi_\rho(t;p,q)$ be the function of form* (2.1). *Then*

$$\begin{aligned}&\left\|\rho(\cdot)-Q_\mu^m(\cdot;\Phi_\rho(t;\cdot,\cdot))\right\|_{C^l(\mu^{-1/2})}\\&\quad\le C\left\|\rho\right\|_{C^{l+\sigma}}\{\mu^{-\sigma}+(C\mu)^{m+n+2}m^{-m/2+l}\}\\&\qquad+e^{C'\mu^2}e^{-C\lambda_0 t}\max_{|\beta|\le m}\sum_{i,j=1}^{N(m)}\left|C_{ij}^\beta\right|\left\|p_i^m\right\|\left\|p_j^m\right\|\end{aligned}$$

for constants C, C' independent of ρ, μ, and m.

Corollary 3.3 implies that we can reconstruct ρ approximately via given $\Phi_\rho(t;\cdot,\cdot)$. Indeed, for any $\epsilon>0$ and compact set D in Ω, there exist μ, m, and t such that $\left\|\rho(\cdot)-Q_\mu^m(\cdot;\Phi_\rho(t;\cdot,\cdot))\right\|_{C^l(D)}<\epsilon$.

Analysis of the estimate in Corollary 3.3 involves an estimate of $\sum_{i,j=1}^{N(m)}|C_{ij}^\beta|\left\|p_i^m\right\|\left\|p_j^m\right\|$. Together with an optimal choice of the parameter μ this gives the final estimate which involves only m and t.

THEOREM 3.4. *Let $\Phi_\rho(t;p,q)$ be the functional defined by* (2.1) *and $Q^m(x;\Phi_\rho)$—the polynomials defined by formula* (3.1) *with $M=\Phi_\rho$ and $\mu=e^{-a}m^{1/2}$. If we choose $a>0$ sufficiently large, then there exist constants $C>0$, $c_1>0$, and $c_2>0$ such that*

$$\begin{aligned}&\left\|\rho(\cdot)-Q^m(\cdot;\Phi_\rho(t;\cdot,\cdot))\right\|_{C^l(e^{a/2}m^{-1/4})}\\&\qquad\le C\{(\left\|\rho\right\|_{C^{l+\sigma}}+1)m^{-\sigma/2}+e^{c_1m-c_2\lambda_0 t}\}.\end{aligned}$$

Proof. We employ the directional moments of order q:

$$X_e^q(x)=\langle x,e\rangle^q,\qquad e\in\mathbf{S}^{n-1},$$

where $\mathbf{S}^{n-1}$ is the unit sphere in $\mathbb{R}^n$ and $\langle\cdot,\cdot\rangle$ stands for the scalar product in $\mathbb{R}^n$. The main idea of the proof is to find a representation

$$x^\alpha=\sum_{\gamma=1}^{P(q)}c_\gamma^\alpha X_{e_\gamma}^q(x)$$

and to estimate $\sum_{\gamma=1}^{P(q)}|c_\gamma^\alpha|$, $|\alpha|=q\le m$.

Consider polynomials of the form $X_{e_1}^{q_1}(x)\ X_{e_2}^{q_2}(x)$, where $\langle e_1,e_2\rangle=0$. Then (for details see [K-S])

$$X_{e_\phi}^q(x)=\sum_{q_1+q_2=q}C_{q_1}^q\cos^q(\phi)\ \tan^{q_1}(\phi)\ X_{e_1}^{q_1}(x)X_{e_2}^{q_2}(x),\tag{3.4}$$

where $e_\phi=e_1\cos(\phi)+e_2\sin(\phi)$ and $C_{q_1}^q$ are the binomial coefficients.

Equation (3.4) with $\phi=\phi_1,\dots,\phi_{q+1}$, where $\tan(\phi_i)\ne\tan(\phi_j)$, $i\ne j$, form a system of linear equations for the unknown $X_{e_1}^{q_1}(x)X_{e_2}^{q_2}(x)$. The corresponding matrix is essentially the Vandermonde matrix for $\tan(\phi_i)$, $i=1,\dots,q+1$. In the following we take $\tan(\phi_i)=1+(i-1)/q$. Then

$$X_{e_1}^{q_1}(x)X_{e_2}^{q_2}(x)=\sum_{i=1}^{q+1}\frac{1}{C_{q_1}^q\cos^q(\phi_i)}\ \frac{\Delta_{q_1,i}}{\Delta}X_{e_{\phi_i}}^q(x),\tag{3.5}$$

where Δ, $\Delta_{q_1,i}$ denote the determinant and (q_1,i) minor of the Vandermonde matrix, correspondingly.

Denote by $\Delta_{q_1}(z)$ the determinant of the Vandermonde matrix with z instead of $\tan(\phi_{q_1})$:

$$\Delta_{q_1}(z) = \sum_{j=1}^{q+1} \Delta_{q_1,j} z^{j-1}.$$

As, on the other hand,

$$\Delta_{q_1}(z) = \Pi_{i<j;i,j\neq q_1}(\tan(\phi_i)-\tan(\phi_j))\Pi_{i<q_1}(z-\tan(\phi_i))\Pi_{i>q_1}(\tan(\phi_i)-z),$$

we obtain that

$$\frac{\Delta_{q_1,j}}{\Delta} = \frac{1}{2\pi i}\int_\Gamma \frac{\Delta_{q_1}(z)}{\Delta}\frac{dz}{z^j}, \tag{3.6}$$

where Γ is, e.g., a circle of the radius 1. However,

$$\frac{\Delta_{q_1}(z)}{\Delta} = \Pi_{i<q_1}\frac{(z-\tan(\phi_i))}{(\tan(\phi_{q_1})-\tan(\phi_i))}\ \Pi_{i>q_1}\frac{(\tan(\phi_i)-z)}{(\tan(\phi_i)-\tan(\phi_{q_1}))}.$$

However, as $\tan(\phi_i) = 1+(i-1)/q$,

$$\Pi_{i\neq q_1}|(\tan(\phi_i)-\tan(\phi_{q_1}))| \geq \left(\frac{q!}{q^q C_{q_1}^q}\right) \geq \frac{C^{-q}}{C_{q_1}^q}.$$

Substitution of this estimate into (3.6) leads to the estimate

$$\left|\frac{\Delta_{q_1,j}}{\Delta}\right| \leq C^q C_{q_1}^q. \tag{3.7}$$

Returning to the estimate for $X_{e_1}^{q_1}(x)\ X_{e_2}^{q_2}(x)$, $q_1+q_2=q$, we use the estimate $\cos(\phi_i)\geq 5^{-1/2}$. Hence (3.5), (3.7) yield that

$$X_{e_1}^{q_1}(x)\ X_{e_2}^{q_2}(x) = \sum_{i=1,\dots,q+1} c_{q_1,i} X_{e_{\phi_i}}^q(x), \quad |c_{q_1,i}|\leq C^q. \tag{3.8}$$

Let us consider $x^\alpha = x_1^{\alpha_1}\cdots x_n^{\alpha_n}$. In the same way as in the proof of Proposition 3 in [K-S], by induction in the number of the variables $x_1,\dots,x_n$, we can show that the formula (3.8) yields the representation

$$x^\alpha = \sum_{\gamma=1}^{P(q)} c_\gamma^\alpha X_{e_\gamma}^q(x), \quad |\alpha| = q,$$

for some $e_\gamma \in S^{n-1}$. From the steps of the induction, it follows that

$$P(q)\leq (1+q)^n, \quad |c_\gamma^\alpha|\leq C^q.$$

Furthermore

$$X_e^q = 2^{-q}(Z_e+\overline{Z}_e)^q = 2^{-q}\sum_{q_1=0}^{q} C_{q_1}^q Z_e^{q_1}\,\overline{Z}_e^{q-q_1},$$

where $Z_e = \langle x, e\rangle + i\langle x, e'\rangle$ for an arbitrary e' such that $\langle e, e'\rangle = 0$. Then

$$x^\alpha = \sum_{\gamma=1}^{P(q)} c_\gamma^\alpha X_{e_\gamma}^q = \sum_{\gamma=1}^{P(q)} c_\gamma^\alpha 2^{-q} \sum_{q_1=0}^{q} C_{q_1}^q Z_{e_\gamma}^{q_1} \overline{Z}_{e_\gamma}^{q-q_1}$$
$$= \sum_{j,k=1}^{N(q)} c_{j,k}^\alpha \, p_j(x)\overline{p}_k(x).$$

Here p_j are harmonic polynomials of the form $p_j(x) = Z_{e_\gamma}^{q_1}(x)$ with $q_1 \le q$ and $p_k(x) = Z_{e_\gamma}^{q-q_1}(x)$. Moreover, as $\sum_{q_1=0}^{q} C_{q_1}^q = 2^q$, we have

$$\sum_{j,k}^{N(q)} \left|c_{j,k}^\alpha\right| = \sum_{\gamma=1}^{P(q)} \left|c_\gamma^\alpha\right| 2^{-q} \sum_{q_1=0}^{q} \left|C_{q_1}^q\right| \le q^n C^q \le (C_n)^q.$$

As $\|p_j\| \, \|p_k\| \le V(\Omega)(1+r_1)^q$, where $V(\Omega)$ is the volume of Ω, the above estimate together with Theorem 3.2 and Corollary 3.3 give rise to the following estimate:

$$\left\|\rho(\cdot) - Q_\mu^m(\cdot; \Phi_\rho(t;\cdot,\cdot))\right\|_{C^l(\mu^{-1/2})} \le C\{\|\rho\|_{C^{l+\sigma}} (\mu^{-\sigma}$$
$$+(C\mu)^{m+n+2} m^{-m/2+l}) + e^{C'\mu^2 - C\lambda_0 t} C^m\}.$$

Thus, inserting $\mu = e^{-a} m^{1/2}$ into the above inequality and taking $a > 0$ large enough, we obtain the estimate in Theorem 3.4. □

Analyzing the proof of Theorem 3.4, we obtain also the following stability estimate which will be used in section 4.

LEMMA 3.5. *Let $M, \widetilde{M}$ be bilinear forms on HP^m. Let $Q_\mu^m(x; M)$ and $Q_\mu^m(x; \widetilde{M})$ be given by formula* (3.1). *Then we have*

$$\left\|Q_\mu^m(\cdot; M) - Q_\mu^m(\cdot; \widetilde{M})\right\|_{C^l(\Omega)} \le C^m e^{c'\mu^2} \left\|M - \widetilde{M}\right\|_m.$$

4. Analytic estimates and stability. In the analysis of section 3 we have not used the fact that $\Phi_\rho(t;\cdot,\cdot)$ is an analytic function when Re $(t) > 1$, which makes it possible to improve the estimates of Corollary 3.3 and Theorem 3.4 and to obtain some further stability results for the considered inverse problem. Namely, by means of the analyticity, we can obtain the required estimates on an interval given in advance (e.g., on $0 < t < 2$).

We start with stability estimates.

LEMMA 4.1. *Let $\rho, \widetilde{\rho}$ satisfy the condition* (1.2) *and*

$$\left\|\Phi_\rho(\cdot; p, q) - \Phi_{\widetilde{\rho}}(\cdot; p, q)\right\|_{C^0(0,2)} \le \epsilon \, \|p\| \, \|q\|, \quad p, q \in HP^m.$$

Then we have

$$|M_\rho(p,q) - M_{\widetilde{\rho}}(p,q)| \le C e^{-C\lambda_0^{1/2} |\lg \epsilon|^{1/2}} ||p|| \, ||q||. \tag{4.1}$$

Proof. Set $z = \frac{t-2}{t}$. Then the map $t \to z$ becomes conformal from the half plane Re $(t) > 1$ onto the unit disk $|z| < 1$. Consider the function

$$f(z) = \Phi_\rho(t; p, q) - \Phi_{\widetilde{\rho}}(t; p, q), \quad z = z(t),$$

which is analytic in the unit disk. Moreover, by Theorem 2.1, we have

$$\begin{aligned} |f(z)| &\le C||p||\ ||q|| \qquad \text{in} \quad |z| \le 1, \\ |f(z)| &\le \epsilon||p||\ ||q|| \qquad \text{on} \quad -1 \le z \le 0. \end{aligned}$$

By the Milloux theorem (see, e.g., [G; Chapter VIII, section 4, Theorem 6]) these estimates imply that when $z = 1 - \zeta$, Im $(z) = 0$,

$$|f(z)| \le C||p||\ ||q||\epsilon^{\zeta/\pi}.$$

Taking $\zeta = 2/t$, we see that

$$|\Phi_\rho(t; p, q) - \Phi_{\tilde\rho}(t; p, q)| \le C||p||\ ||q||\epsilon^{2/\pi t}.$$

This estimate together with Theorem 2.1 (iii), where $t = (\frac{2|\lg\epsilon|}{C\pi\lambda_0})^{1/2}$, proves the lemma.

The proof is complete. □

Lemma 4.1 together with Theorem 3.2(ii) and Lemma 3.5 leads to the following stability result.

THEOREM 4.2. *Let $\rho, \widetilde\rho$ satisfy the condition* (1.2) *and*

$$||\Phi_\rho(\cdot; p, q) - \Phi_{\tilde\rho}(\cdot; p, q)||_{C^0(0,2)} \le \epsilon||p||||q||, \quad p, q \in HP^m.$$

Then we have

$$\begin{aligned} \left\|\rho - \widetilde\rho\right\|_{C^l(\mu^{-1/2})} \le C[(\left\|\rho\right\|_{C^{l+\sigma}} + \left\|\tilde\rho\right\|_{C^{l+\sigma}})\{\mu^{-\sigma} + (C\mu)^{m+n+2}m^{-m/2+l}\} \\ + C^m e^{C\mu^2} e^{-C\lambda_0^{1/2}|\lg\epsilon|^{1/2}}]. \end{aligned}$$

Proof. Obviously

$$\begin{aligned} \left\|\rho - \tilde\rho\right\|_{C^l(\mu^{-1/2})} \le \left\|\rho - Q_\mu^m(x, M_\rho)\right\|_{C^l(\mu^{-1/2})} \\ + \left\|\tilde\rho - Q_\mu^m(x, M_{\tilde\rho})\right\|_{C^l(\mu^{-1/2})} \\ + \left\|Q_\mu^m(x, M_\rho) - Q_\mu^m(x, M_{\tilde\rho})\right\|_{C^l(\mu^{-1/2})}. \end{aligned} \tag{4.2}$$

The first two terms in the right-hand side of (4.2) may be estimated by means of (3.3). To estimate the third term we use the relation (4.1) together with Lemma 3.5 with $M_\rho, M_{\tilde\rho}$ instead of $M, \widetilde M$. Therefore the theorem is obtained. The proof is now complete. □

Remark 1. If we take $\mu = e^{-a}m^{1/2}$, where a is a sufficiently large positive number which depends upon Ω, ρ_1, ρ_2, and l, the above estimate may be simplified in the following way:

$$\begin{aligned} \left\|\rho - \widetilde\rho\right\|_{C^l(e^{a/2}m^{-1/4})} \le C\{(\left\|\rho\right\|_{C^{l+\sigma}} + \left\|\tilde\rho\right\|_{C^{l+\sigma}})m^{-\sigma/2} \\ + \exp(C_1 m - C_2\lambda_0^{1/2}|\lg\epsilon|^{1/2})\}. \end{aligned}$$

Remark 2. From the proof of Theorem 4.2, we can also obtain the following conditional stability result: There exist constants a, C, and C_1 such that the inequality

$$\left\|\rho - \widetilde\rho\right\|_{C^l(e^{a/2}m^{-1/4})} \le CEm^{-\sigma/2} + e^{C_1 m}\epsilon$$

holds for any ρ, $\widetilde{\rho} \in C^{l+\sigma}(\overline{\Omega})$, satisfying $\|\rho\|_{C^{l+\sigma}(\overline{\Omega})} \leq E$ and $\|\widetilde{\rho}\|_{C^{l+\sigma}(\overline{\Omega})} \leq E$ if we have

$$|M_\rho(p,q) - M_{\tilde{\rho}}(p,q)| \leq \epsilon \|p\| \|q\| \qquad \text{for any } p, q \in HP^\infty.$$

Our next goal is to improve the reconstruction procedure described in section 3 for the case when $\Phi_\rho(t;\cdot,\cdot)$ is known with some error.

Let $\Psi_\epsilon(t;p,q)$ $(p,q \in HP^m)$ satisfy the following estimates on $0 \leq t \leq 2$:

$$\|\Phi_\rho(\cdot;p,q) - \Psi_\epsilon(\cdot;p,q)\|_{C^0(0,2)} \leq \epsilon \|p\| \|q\|. \tag{4.3}$$

Set

$$w(t) = \frac{\{1-(t-1)^2\}^{1/2} - i}{\{1-(t-1)^2\}^{1/2} + i}.$$

Then $w(t)$ becomes a conformal map of the half plane Re $(t) > 1$ with a slit along the interval $(1,2)$ onto the unit disk $|w| < 1$. The slit is transformed onto the left semicircle $|w| = 1, \text{Re}\,(w) < 0$, and the line Re $(t) = 1$ onto the right semicircle $|w| = 1, \text{Re}\,(w) > 0$. The function $f_\epsilon(w;p,q) = \Psi_\epsilon(t;p,q)$ $(t \in (1,2),\ w = w(t))$ is defined on the left semicircle and satisfies

$$|f(w) - f_\epsilon(w)| \leq \epsilon \|p\| \|q\| \qquad \text{on } |w| = 1,\ \text{Re}\,(w) < 0, \tag{4.4}$$

where $f(w;p,q) = \Phi_\rho(t;p,q)$ $(t \in (1,2),\ w = w(t))$.

For the quasi-analytic continuation of $f_\epsilon(w)$ we use the construction suggested in [L] which is based upon the Carleman lemma [Car]. Let

$$\sigma_\epsilon = \frac{1}{t^2} \log \frac{Ct}{\epsilon},$$

with some $t \geq 2$. We define $\tilde{f}_\epsilon^t(z)$ by the following formula:

$$\tilde{f}_\epsilon^t(z) = \frac{e^{-\sigma_\epsilon}}{2\pi i} \int_{\frac{\pi}{2}}^{\frac{3\pi}{2}} f_\epsilon(w) \exp\left\{\sigma_\epsilon \left(\frac{w-1}{z-1}\right)^2\right\} \frac{iwd\phi}{z-w}, \quad w = e^{i\phi}. \tag{4.5}$$

THEOREM 4.3. *Let $\tilde{f}_\epsilon^t(z)$ be given by* (4.5)*, where f_ϵ satisfies* (4.4)*. Then for $t > 2$ large enough, we have*

$$|M_\rho(p,q) - \tilde{f}_\epsilon^t(1-2/t)| \leq C \|p\| \|q\| (t\epsilon^{1/t^2} + e^{-C\lambda_0 t}). \tag{4.6}$$

Proof. By $\mathcal{O}$ (see Figure 1), we denote the domain obtained as the intersection of the unit disk and the sector of the angle $\pi/2$ with its vertex in the point $(1,0)$, which is symmetric with respect to the real axis.

Since the function $f(w;p,q)$ is continued analytically in the unit disk, by the Cauchy formula for the holomorphic function $z \to f(z)\exp\left(\frac{\sigma_\epsilon(z-1)^2t^2}{4}\right)$, we have

$$f(1-2/t) = \frac{e^{-\sigma_\epsilon}}{2\pi i} \int_{\partial\mathcal{O}} f(z) \exp\left(\frac{\sigma_\epsilon(z-1)^2t^2}{4}\right) \frac{dz}{1-2/t-z},$$

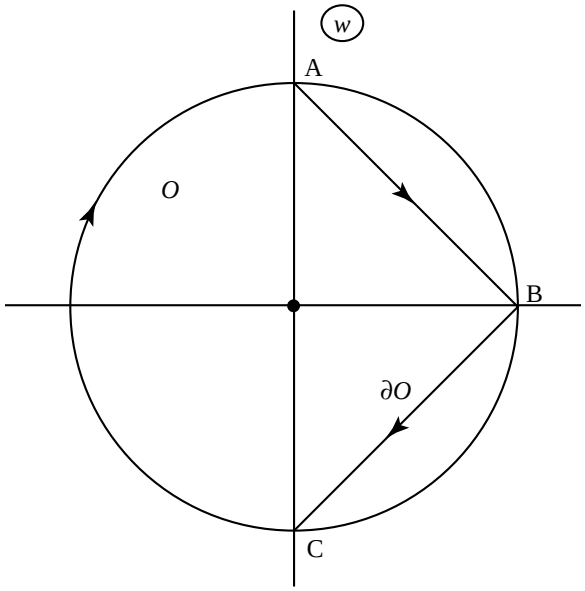


FIG. 1.

where the contour $\partial\mathcal{O}$ consists of the left semicircle (in Re $(z) < 0$) and the broken line ABC (see Figure 1). Hence

$$
\begin{aligned}
&|\tilde{f}_\epsilon^t(1-2/t) - f(1-2/t)| \\
&\leq \frac{e^{-\sigma_\epsilon}}{2\pi} \int_{\frac{\pi}{2}}^{\frac{3\pi}{2}} |\tilde{f}_\epsilon(z) - f(z)| \exp\left(\frac{\sigma_\epsilon t^2}{4} \mathrm{Re}\ (z-1)^2\right) \left|\frac{dz}{1-2/t-z}\right| \\
&+ \frac{e^{-\sigma_\epsilon}}{2\pi} \int_{ABC} |f(z)| \exp\left(\frac{\sigma_\epsilon t^2}{4} \mathrm{Re}\ (z-1)^2\right) \left|\frac{dz}{1-2/t-z}\right| \\
&= I_1 + I_2,
\end{aligned} \tag{4.7}
$$

where $z = e^{i\phi}, \phi \in (\frac{\pi}{2}, \frac{3\pi}{2})$, in the first integral I_1.

For $z = e^{i\phi}$, $\phi \in (\frac{\pi}{2}, \frac{3\pi}{2})$, we have Re $(z-1)^2 \leq 2$ and $|1-2/t-z| \geq 1$. Hence the inequality (4.4) yields the following estimate for I_1:

$$
I_1 \leq C\epsilon \left\|p\right\| \left\|q\right\| e^{\sigma_\epsilon(t^2-1)} \leq Ct\epsilon^{1/t^2} \left\|p\right\| \left\|q\right\|, \tag{4.8}
$$

where the last estimate follows from the definition of σ_ϵ.

To estimate I_2, we use the fact that Re $\{(z-1)^2\} = 0$ and $|1-2/t-z| \geq \frac{\sqrt{2}}{t}$ on ABC. Taking into account the estimate (i) of Theorem 2.1 and $|\Phi_\rho(t,p,q)| \leq C \left\|p\right\| \left\|q\right\|$ for Re $t \geq 1$, we see that

$$
I_2 \leq Ct \left\|p\right\| \left\|q\right\| e^{-\sigma_\epsilon} \leq Ct\epsilon^{1/t^2} \left\|p\right\| \left\|q\right\|. \tag{4.9}
$$

Since we clearly have $|w(t) - (1-2/t)| \leq C/t^2$, there is a constant $C > 0$ such that $\left|t - \tilde{t}\right| \leq C$ holds for $t \geq 1$ and $\tilde{t}$ satisfying $1 - 2/t = w(\tilde{t})$. Hence, taking into

account Theorem 2.1 (iii), we see that

$$\left|f(1-2/t)-M_\rho\right| = \left|\Phi_\rho(\tilde{t};p,q)-M_\rho(p,q)\right| \le Ce^{-C'\lambda_0 t}\left\|p\right\|\left\|q\right\|. \tag{4.10}$$

The inequality (4.6) follows from (4.7)–(4.10).

The proof is complete. □

Remark. The consideration leading to (4.9) is a special case of the results obtained by Lavrent'ev (see [L, Chapter I, Proof III]).

For any bilinear form $\Psi_\epsilon(t;p,q)$ from $HP^m \times HP^m$ to $C^0[0,2]$, we set

$$M_\rho^\epsilon(p,q) = \tilde{f}_\epsilon^t(1-2/t), \qquad t = |\log\epsilon|^{1/3}\lambda_0^{-1/3}, \tag{4.11}$$

where $\tilde{f}_\epsilon^t$ is the function in (4.5) with $f_\epsilon(w) = \Psi_\epsilon(t;p,q)$, $w = w(t)$. Then Theorem 4.3 implies that

$$\left|M_\rho(p,q) - M_\rho^\epsilon(p,q)\right| \le C|\log\epsilon|^{1/3}e^{-C'\lambda_0^{2/3}|\log\epsilon|^{1/3}}\left\|p\right\|\left\|q\right\|.$$

By the above estimate and the estimates (3.3), (3.2), and proof of Theorem 3.4, we come to the following.

THEOREM 4.4. *Let $\Psi_\epsilon : HP^m \times HP^m \to C^0(0,2)$ be a bilinear form which is ϵ-close to Φ_ρ in the sense of* (4.3). *For this Ψ_ϵ, we define M_ρ^ϵ by* (4.11). *Then we have*

$$\begin{aligned}\left\|\rho(\cdot) - Q_\mu^m(\cdot;M_\rho^\epsilon)\right\|_{C^l(\mu^{-1/2})} \le C[&\left\|\rho\right\|_{C^{l+\sigma}}(\mu^{-\sigma}+(C\mu)^{m+n+2}m^{-m/2+l})\\ &+|\log\epsilon|^{1/3}\mu^n e^{C''\mu^2}C^m e^{-C'\lambda_0^{2/3}|\log\epsilon|^{1/3}}].\end{aligned}$$

In particular, for $\mu = e^{-a}m^{1/2}$ with $a > 0$ large enough, we have

$$\begin{aligned}&\left\|\rho(\cdot) - Q^m_{e^{-a}m^{1/2}}(\cdot;M_\rho^\epsilon)\right\|_{C^l(e^{a/2}m^{-1/4})}\\ &\le C(1+\left\|\rho\right\|_{C^{l+\sigma}})(m^{-\sigma/2}+\exp(Cm - C'\lambda_0^{2/3}|\log\epsilon|^{1/3})).\end{aligned}$$

Acknowledgments. The authors would like to thank Professor G. Nakamura, Professor E. Somersalo, and Professor K. Peat for their interest in this paper.

REFERENCES

[Cal] A. CALDERON, *On an inverse boundary value problem*, in Seminar on Numerical Analysis and Its Applications to Continuum Physics, Soc. Brasil. Mat., Rio de Janeiro, 1980, pp. 65–73.

[Car] T. CARLEMAN, *Les Fonctions Quasi-Analytiques*, Gauthier-Villars, Paris, 1926.

[G] G. M. GOLUZIN, *Geometric Theory of Functions of a Complex Variable*, Transl. Math. Monogr. 26, AMS, Providence, RI, 1969.

[Is] V. ISAKOV, *Inverse Problems for Partial Differential Equations*, Appl. Math. Sci. 127, Springer, New York, 1998, p. 284.

[K-K-S] M. KAWASHITA, Y. KURYLEV, AND H. SOGA, *A moment method on inverse problems for the heat equation*, in Proceedings of the Japan-Korea Joint Scientific Seminar on Inverse Problems and Related Topics, Res. Notes Math. 419, Chapman & Hall/CRC, Boca Raton, FL, 2000.

[K-P] Y. KURYLEV AND K. S. PEAT, *Hausdorff moments in two-dimensional inverse acoustic problem*, Inverse Problems, 13 (1997), pp. 1363–1377.

[K-S] Y. Kurylev and A. Starkov, *Directional moments in the acoustic inverse problem*, in Inverse Problems in Wave Propagation, G. Chavent, et al., eds., IMA Vol. Math. Appl. 90, Springer, New York, 1997, pp. 295–324.

[L] M. M. Lavrent'ev, *Cauchy problem for the Laplace equation*, Izv. Akad. Nauk SSSR Ser. Matem., 20 (1956), pp. 819–842 (in Russian).

[S-U] J. Sylvester and G. Uhlmann, *A global uniqueness theorem for an inverse boundary value problem*, Ann. Math., 125 (1987), pp. 153–169.

SIAM J. MATH. ANAL.
Vol. 32, No. 3, pp. 538–554

EXPLICIT COMPUTATION OF ORTHONORMAL SYMMETRIZED HARMONICS WITH APPLICATION TO THE IDENTITY REPRESENTATION OF THE ICOSAHEDRAL GROUP*

YIBIN ZHENG† AND PETER C. DOERSCHUK‡

Abstract. A novel method to explicitly compute orthonormal symmetrized harmonics is presented and the method is applied to the identity representation of the icosahedral group. Spherical viruses have icosahedral symmetry and the motivating application is the parametric representation of spherical viruses for use in inverse problems based on x-ray scattering data and cryoelectron microscopy images. The symmetrized harmonics are computed in the form of linear combinations of spherical harmonics of one order and therefore have simple rotational properties which is valuable in the electron microscopy application. The method is based on equating the expansions of a symmetrized delta function in spherical and in symmetrized harmonics from which bilinear equations for the weights in the linear combinations can be derived. The explicit character of the calculation is reflected in the fact that both explicit expressions and an efficient recursive algorithm are derived for computing the weights in the linear combinations.

Key words. symmetric harmonics, icosahedral harmonics

AMS subject classifications. 33C55, 20C40, 20H15, 20G45, 92B05

PII. S0036141098341770

1. Introduction.

An important problem in biophysics is the determination of the three-dimensional distribution of electron density in so-called "spherical" viruses [1] from x-ray scattering and electron microscopy data [2]. This is a large class of viruses, including both viruses of plants and animals, where all viral particles of a particular viral type are identical, each viral type has a particle diameter of 10^2–10^3Å, and each viral particle has all the symmetries of the icosahedron. Two approaches to analyzing such data are to represent the electron density either as a truncated orthonormal expansion [3, 4] or as a piecewise-constant function with icosahedrally symmetric boundaries that are described using truncated orthonormal expansions [5] and then solve a nonlinear least squares problem for the coefficients in the expansion. Because the viral particles are roughly spherical in shape and the icosahedral symmetry is a rotational symmetry, it is natural to use spherical coordinates and express the basis functions in the orthonormal expansion of the electron density as products of functions on the sphere and radial functions. Similarly, for the approach based on the piecewise-constant function, it is natural to describe the boundary by its radius from the origin as a function of the angles, in which case the basis functions in the expansion are functions on the sphere.

*Received by the editors July 13, 1998; accepted for publication (in revised form) January 21, 2000; published electronically September 15, 2000. This work was partially supported by NSF grants BIR-9513594 and DBI-9630497.

http://www.siam.org/journals/sima/32-3/34177.html

†GE Corporate Research and Development, KWC-605, One Research Circle, Niskayuna, NY 12309 (zheng@crd.ge.com). Part of this work was done while this author was affiliated with Purdue University.

‡School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907-1285 (doerschu@ecn.purdue.edu).

In this paper we describe a new method for explicitly[1] computing sets of functions that are representations of rotational groups in three dimensions and demonstrate the method on the identity representation of the icosahedral group, i.e., we explicitly compute a complete orthonormal basis for icosahedrally symmetric functions on the sphere. We call the functions in this basis "icosahedrally symmetric basis functions" (abbreviated ISBFs) and are not yet specific about which basis we will explicitly compute (see section 2). Using the ISBFs in the orthonormal expansions involved in the biophysics problems of the previous paragraph is much superior to the natural alternative of using spherical harmonics (denoted by $Y_{l,m}(\theta,\phi)$, where, here and elsewhere, (θ,ϕ) are spherical coordinates). For example: (1) The constraint that the particle has icosahedral symmetry is built into the functions rather than having to be added as a constraint in the nonlinear least squares problem. (2) There are many fewer ISBFs than $Y_{l,m}$ functions so many fewer coefficients have to be determined by nonlinear least squares in order to determine the electron density at a given level of resolution. (3) By incorporating the icosahedral symmetry directly in the mathematical description of the electron density by use of the ISBFs, we remove certain nonuniqueness problems in the nonlinear least squares problems.

There has been extensive work on ISBFs [7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19], which are basis functions for the identity representation of the icosahedral group (the only representation needed in our application), and also on the more general problem of basis functions for all five irreducible representations of the icosahedral group [9, 17]. Sometimes ISBFs are called "icosahedral harmonics" [12, 20, 14] but, in analogy with spherical harmonics terminology, we reserve "icosahedral harmonics" for a collection of basis functions for all five irreducible representations. In the majority of previous work, ISBFs are described as linear combinations of spherical harmonics of fixed order which leads to simple rotational properties which are important for our electron microscopy applications. In this framework, the only task is to determine the coefficients of the linear combination. In a minority of previous work (e.g., [14]), ISBFs are described as polynomials in the rectangular coordinates.

Although extensive work has been done, existing results are limited in two aspects that are important for our application: (1) Explicit expressions in terms of standard operations ($+$, $-$, $\times$, $\div$, and complex exponentiation) for ISBFs of arbitrary order are not provided. (2) The algorithms provided to derive an ISBF for some particular order are laborious, especially for orders greater than 29 when there can be two or more ISBFs of a single order. Reflecting these limitations, the most extensive tables of which we are aware [12, 18] tabulate ISBFs only up to order 30 (and in fact only one of two functions of order 30 is tabulated) or 44, respectively, while in a medium resolution x-ray diffraction interpolation problem we have required functions of order roughly 85. In the previous work [18] most closely related to the present paper, the derivation is unnecessarily complicated and not rigorous due to the "Q" operator in [18], implementation of the resulting algorithm requires symbolic derivatives in

[1] By "explicit" we mean relationships from which formulas such as those given in section 7 can be computed. This is a weaker notion of "explicit" than is standard in, for example, spherical harmonics, where [6, Eqs. (3.53) and (3.50)]

$$Y_{l,m}(\theta,\phi) = \sqrt{\frac{2l+1}{4\pi}\frac{(l-m)!}{(l+m)!}}P_{l,m}(\cos\theta)\exp(im\phi),$$

$$P_{l,m}(x) = \frac{(-1)^m}{2^l l!}(1-x^2)^{m/2}\frac{\mathrm{d}^{l+m}}{\mathrm{d}x^{l+m}}(x^2-1)^l.$$

order to compute the "$c_{p,q}$" numbers in [18] (numerical approximations of the derivatives are essentially impossible because determination of an lth order ISBF requires derivatives through lth order), the resulting algorithm in [18] is very slow compared to the algorithm of this paper, and the matrix-factorization implications of (15) in [18] are not appreciated or exploited. In another related work [19], the authors base their approach on projection operators applied to spherical harmonics (as in equation (1) of the present paper) followed by Gram–Schmidt orthogonalization and do all of the necessary integrals numerically to machine precision by noting that the integrands of interest are polynomials and that, therefore, suitably high-order Gaussian quadrature formulas can compute the integrals to machine precision. These machine precision results are then represented using square roots and rational numbers, and tables to order 15 are provided. In contrast with the approach of [19], we apply projection operators to delta functions rather than spherical harmonics, expand the result in spherical harmonics and in ISBFs, and by equating the two expansions derive a bilinear system of equations from which the weights in the expansion of an ISBF in terms of spherical harmonics can be derived symbolically rather than numerically to machine precision. In summary, in this paper we remove the limitations indicated above and the resulting algorithm is highly suitable for computer implementation in either numerical or symbolic programming languages; software is available from the authors.

2. Approach. Our goal is to determine ISBFs such that (1) each function is real valued, (2) the set of functions are a complete orthonormal basis for smooth icosahedrally symmetric functions on the sphere, and (3) each function is a linear combination of $Y_{l,m}$ for some fixed l. Let $\int \mathrm{d}\Omega$ mean $\int_{\theta=0}^{\pi}\int_{\phi=0}^{2\pi}\sin(\theta)\mathrm{d}\theta\mathrm{d}\phi$. Then the orthonormality referred to in the second property is $\int I_\alpha^*(\theta,\phi)I_{\alpha'}(\theta,\phi)\mathrm{d}\Omega = \delta_{\alpha,\alpha'}$, where I_α denotes an ISBF from a not-yet-specified basis. Since rotation of $Y_{l,m}$ gives a function that is a linear combination of $\{Y_{l,m'} : m' = -l,\ldots,+l\}$ [21], the third property assures that the ISBFs will have simple properties under rotations, which is important in electron microscopy [2].

Goals (1)–(3) of the previous paragraph do not uniquely define the ISBFs when there are two or more ISBFs (denoted by $I_{\alpha_1},\ldots,I_{\alpha_p}$) in the subspace $\mathcal{S}_l$ spanned by $\{Y_{l,m} : m = -l,\ldots,+l\}$. In particular, if $O \in \mathcal{R}^{p\times p}$ is an orthonormal matrix, then the p functions $(I_{\alpha_1},\ldots,I_{\alpha_p})^T$ could be replaced by the p functions $O(I_{\alpha_1},\ldots,I_{\alpha_p})^T$ and still satisfy goals (1)–(3). A method to chose the basis in the subspace $\mathcal{S}_l$ so that the basis has meaningful representation-theoretic or spectral-theoretic properties is an open question. In this paper (see section 6) the basis is chosen so that the matrix of expansion coefficients, expanding ISBFs in terms of spherical harmonics, is triangular. This choice of basis minimizes the number of terms when computing ISBFs from spherical harmonics. Except for sections 6–8, all results in the paper are true for any basis satisfying goals (1)–(3).

A standard group-theoretic approach to determine the ISBFs is to apply projection operators [22, pp. 92–94] to the spherical harmonics. For the identity representation of a group, the projection operator has a simple form and a candidate ISBF, that is, the projection operator applied to the (l,m)th spherical harmonic, is

$$J_{l,m}(\theta,\phi) = \frac{1}{g}\sum_{k=0}^{g-1} P(T_k)Y_{l,m}(\theta,\phi), \tag{1}$$

where $g = 60$ is the order of the icosahedral group, T_k is the kth rotation of the

icosahedral group, and the scalar transformation operator $P(T)$ applied to a function $\psi(\mathbf{r})$ is defined by $P(T)\psi(\mathbf{r}) = \psi(T^{-1}\mathbf{r})$. (We are using the notation of [22].) While this method appears to be direct, it has some serious difficulties: First,

$$P(T_k)Y_{l,m}(\theta,\phi) = \sum_{m'=-l}^{+l} D_{l,m,m'}(T_k)Y_{l,m'}(\theta,\phi),$$

where the $D_{l,m,m'}(T_k)$ are the complicated Wigner's D coefficients [21], so it is difficult to perform the sum of (1) analytically for general l and m. Second, for a fixed l, Laporte's results (see Theorem 3.1) state that there are only $N_l \leq 2l+1$ linearly independent ISBFs that can be constructed from $\{Y_{l,m} : m = -l, \ldots, +l\}$) while (1) will generate $2l+1$ candidates. Therefore, N_l functions must be chosen from among the $2l+1$ candidates. Furthermore, no set of N_l functions from among the candidates is guaranteed to be orthonormal, so a set of N_l linearly independent functions must then be orthogonalized by the Gram–Schmidt procedure. This orthogonalization is also difficult to perform analytically for general l and m. In summary, it is difficult to derive, by way of (1), expressions for an orthonormal set of ISBFs that are explicit functions of the indices.

Our approach is also based on projections. However, rather than projecting a spherical harmonic, as in (1), we project a delta function located at spherical coordinates (θ_0, ϕ_0), i.e., $\delta(\cos\theta - \cos\theta_0)\delta(\phi - \phi_0)$. The result of the projection is a symmetrized delta function denoted by $\Delta(\theta_0, \phi_0; \theta, \phi)$:

$$\Delta(\theta_0,\phi_0;\theta,\phi) = \frac{1}{g}\sum_{k=0}^{g-1} P(T_k)[\delta(\cos\theta - \cos\theta_0)\delta(\phi - \phi_0)].$$

This projection is easy to compute because the result of applying a rotation to a delta function is just another delta function at different coordinates: $P(T_k)[\delta(\cos\theta - \cos\theta_0)\delta(\phi - \phi_0)] = \delta(\cos\theta - \cos\theta_k)\delta(\phi - \phi_k)$. Furthermore, it is straightforward to expand the symmetrized delta function $\Delta(\theta_0, \phi_0; \theta, \phi)$ as a weighted sum of spherical harmonics:

$$\Delta(\theta_0,\phi_0;\theta,\phi) = \sum_{l=0}^{\infty}\sum_{m=-l}^{+l} w_{l,m}(\theta_0,\phi_0)Y_{l,m}(\theta,\phi),$$

specifically,

$$\Delta(\theta_0,\phi_0;\theta,\phi) = \frac{1}{g}\sum_{k=0}^{g-1}\sum_{l=0}^{\infty}\sum_{m=-l}^{+l} Y^*_{l,m}(\theta_k,\phi_k)Y_{l,m}(\theta,\phi). \tag{2}$$

In addition, because the ISBFs are a complete orthonormal fixed basis for the subspace of totally symmetric functions, we know the expansion of $\Delta(\theta_0, \phi_0; \theta, \phi)$ as a weighted sum of ISBFs:

$$\Delta(\theta_0,\phi_0;\theta,\phi) = \sum_{\alpha} I^*_{\alpha}(\theta_0,\phi_0)I_{\alpha}(\theta,\phi). \tag{3}$$

In order to assure that each ISBF is a linear combination of $Y_{l,m}$ for fixed l we constrain the ISBF, denoted by $I_{l,n}$, to have the form

$$I_{l,n}(\theta,\phi) = \sum_{m=-l}^{+l} b_{l,n,m}Y_{l,m}(\theta,\phi), \tag{4}$$

where $l \in \{0, 1, \dots\}$, $n \in \{0, 1, \dots, N_l - 1\}$ (see Theorem 3.1 for the value of N_l), and the weights $b_{l,n,m}$ are unknown and are in fact the goal of these calculations. It is the matrix constructed from $b_{l,n,m}$ (l fixed) that will be made triangular in section 6, thereby selecting a particular orthonormal basis as described above. Finally, by equating (2) and (3) and using (4), we can derive nonlinear equations for the weights $b_{l,n,m}$ and these nonlinear equations can be solved recursively to give explicit formulas for the $b_{l,n,m}$.

3. Preliminaries. For the spherical harmonics $Y_{l,m}$ we use the conventions of [6] and exploit the standard result [6, Eq. (3.53)] that $Y_{l,m}(\theta, \phi) = N_{l,m} P_{l,m}(\cos\theta) e^{im\phi}$, where $P_{l,m}(x)$ are the associated Legendre functions [6, Eq. (3.49)] and

$$N_{l,m} = \sqrt{\frac{2l+1}{4\pi} \frac{(l-m)!}{(l+m)!}}.$$

Laporte [14] proves the following result regarding N_l.

THEOREM 3.1 (Laporte [14]). *For l even, the number N_l (denoted by $N_l^{(\mathrm{e})}$) satisfies the relationship*

$$\frac{1}{(1-x^6)(1-x^{10})} = \sum_{l=0}^{\infty} N_l^{(\mathrm{e})} x^l$$

while for l odd, the number N_l (denoted by $N_l^{(\mathrm{o})}$) is

$$N_l^{(\mathrm{o})} = \begin{cases} N_{l-15}^{(\mathrm{e})}, & l \geq 15, \\ 0, & 0 \leq l < 15. \end{cases}$$

For our concrete calculations, we choose the coordinate system used by Altmann [7] and Laporte [14] in which the z axis passes through two opposite vertices and the xz plane includes one edge of the icosahedron.

4. The bilinear equations for $b_{l,n,m}$. The first proposition about the $b_{l,n,m}$ coefficients can be determined simply from the choice that $I_{l,n}$ are real and $Y_{l,-m}(\theta, \phi) = (-1)^m Y_{l,m}^*(\theta, \phi)$ [6, Eq. (3.54)].

PROPOSITION 4.1. *For each $l = 0, 1, \dots$, $n = 0, \dots, N_l - 1$, and $m = -l, \dots, +l$,*

$$b_{l,n,m} = (-1)^m b_{l,n,-m}^*.$$

The second proposition, based on the orthonormality of the $Y_{l,m}$, relates the orthonormality of the $b_{l,n,m}$ coefficients to the orthonormality of the $I_{l,n}$.

PROPOSITION 4.2. *$I_{l,n}$ and $I_{l',n'}$ ($l \neq l'$; $l, l' = 0, 1, \dots$; $n = 0, \dots, N_l - 1$; $n' = 0, \dots, N_{l'} - 1$) are orthonormal for any choice of $b_{l,n,m}$. For fixed $l = 0, 1, \dots$ the $I_{l,n}$ ($n = 0, \dots, N_l - 1$) are orthonormal if and only if*

$$\sum_{m=-l}^{+l} b_{l,n,m} b_{l,n',m}^* = \delta_{n,n'}.$$

Let (θ_0, ϕ_0) be the (arbitrary) spherical coordinates of a delta function within the first asymmetric unit. Let $\{(\theta_k, \phi_k) : k = 1, 2, \dots, 59\}$ be spherical coordinates of delta functions in the remaining 59 asymmetric units generated by applying rotations

in the icosahedral group. The locations of these additional 59 delta functions are given by Proposition 4.3 below.

PROPOSITION 4.3. *As a function of the parameters θ_0 and ϕ_0, the 60 symmetry-related positions on the unit sphere are*

$$\begin{aligned}
&\{(\theta_k,\phi_k) : k = 0,1,\ldots,59\}\\
&\quad = \{(\theta_0,\phi_k) : k = 0,1,\ldots,4\} \bigcup \left(\bigcup_{n=0}^{4} \left\{ \left(\gamma_n, \alpha_n + k\frac{2\pi}{5} \right) : k = 0,1,\ldots,4 \right\} \right)\\
&\qquad \bigcup \left(\bigcup_{n=0}^{4} \left\{ \left(\pi - \gamma_n, \pi - \alpha_n + k\frac{2\pi}{5} \right) : k = 0,1,\ldots,4 \right\} \right)\\
&\qquad \bigcup \{(\pi - \theta_0, \pi - \phi_k) : k = 0,1,\ldots,4\},
\end{aligned}$$

where ϕ_k, γ_k, and α_k $(k = 0,1,\ldots,4)$ are related to θ_0 and ϕ_0 by

$$\begin{aligned}
\phi_k &= \phi_0 + k\frac{2\pi}{5},\\
\cos\gamma_k &= \frac{1}{\sqrt{5}}(\cos\theta_0 + 2\sin\theta_0\cos\phi_k),\\
\cos\alpha_k &= \frac{2 - \sin\theta_0\cos\phi_k}{\sqrt{5 - (\cos\theta_0 + 2\sin\theta_0\cos\phi_k)^2}}.
\end{aligned}$$

The following proposition is used in the simplification of the the bilinear equation determining the $b_{l,n,m}$ coefficients.

PROPOSITION 4.4. *For any θ_0 and ϕ_0,*

$$\sum_{k=0}^{59} Y_{l,m}(\theta_k,\phi_k) = \begin{cases} 5N_{l,m}\Big[P_{l,m}(\cos\theta_0)\left(e^{im\phi_0} + (-1)^l e^{-im\phi_0}\right) \\ \quad + \sum_{k=0}^{4} P_{l,m}(\cos\gamma_k)\left(e^{im\alpha_k} + (-1)^l e^{-im\alpha_k}\right)\Big], & m = 5\mu \text{ with } \mu \in \mathcal{Z},\\ 0 & \text{otherwise,} \end{cases}$$

where $\mathcal{Z}$ are the integers.

Equate the expressions for $\Delta(\theta_0,\phi_0;\theta,\phi)$ in terms of spherical harmonics (2) and ISBFs (3) to find that

$$\sum_{l=0}^{\infty}\sum_{n=0}^{N_l-1} I_{l,n}(\theta_0,\phi_0) I_{l,n}(\theta,\phi) = \frac{1}{60}\sum_{k=0}^{59}\sum_{l=0}^{\infty}\sum_{m=-l}^{+l} Y^*_{l,m}(\theta_k,\phi_k) Y_{l,m}(\theta,\phi). \tag{5}$$

Replace $I_{l,n}(\theta,\phi)$ by its expansion in terms of $Y_{l,m}(\theta,\phi)$ (4), multiply by $Y^*_{l',m'}(\theta,\phi)$, integrate over solid angles in θ and ϕ, and use the orthonormality of the spherical harmonics to obtain (after renaming the indices $l' \to l$, $m' \to m$) one form (Proposition 4.5, equation (6)) of the fundamental equation for determining the $b_{l,n,m}$ coefficients. Use Proposition 4.4 in (6) or (4) in (6) to obtain two alternative forms ((7) and (8), respectively). The results are summarized in the following proposition.

PROPOSITION 4.5. *The $b_{l,n,m}$ ($l = 0, 1, \ldots$; $n = 0, \ldots, N_l - 1$; $m = -l, \ldots, +l$) coefficients satisfy each of the following equivalent relationships for arbitrary θ_0 and ϕ_0:*

$$\sum_{n=0}^{N_l-1} b_{l,n,m} I_{l,n}(\theta_0, \phi_0) = \frac{1}{60} \sum_{k=0}^{59} Y_{l,m}^*(\theta_k, \phi_k), \tag{6}$$

$$\sum_{n=0}^{N_l-1} b_{l,n,m} I_{l,n}(\theta_0, \phi_0) = \begin{cases} \frac{1}{12} N_{l,m} \Big[P_{l,m}(\cos\theta_0) \left(e^{im\phi_0} + (-1)^l e^{-im\phi_0} \right) \\ \quad + \sum_{k=0}^{4} P_{l,m}(\cos\gamma_k) \left(e^{im\alpha_k} + (-1)^l e^{-im\alpha_k} \right) \Big]^*, & m = 5\mu \text{ with } \mu \in \mathcal{Z}, \\ 0 & \text{otherwise}, \end{cases} \tag{7}$$

$$\sum_{n=0}^{N_l-1} \sum_{m'=-l}^{+l} b_{l,n,m'} b_{l,n,m} Y_{l,m'}(\theta_0, \phi_0) = \frac{1}{60} \sum_{k=0}^{59} Y_{l,m}(\theta_k, \phi_k) \tag{8}$$

for any $l = 0, 1, \ldots$ and $m = -l, \ldots, +l$.

Notice that there is no coupling between different values of l in (6), (7), and (8). From (7) we immediately obtain the following properties of the $b_{l,n,m}$ coefficients.

PROPOSITION 4.6.

1. *If $m \neq 5\mu$ with $\mu \in \mathcal{Z}$, then $b_{l,n,m} = 0$.*
2. *For l even, $b_{l,n,m}$ is real. For l odd, $b_{l,n,m}$ is imaginary.*
3. *$b_{l,n,m} = b_{l,n,-m}(-1)^{l+m}$.*
4. *For l odd, $b_{l,n,0} = 0$.*

Using these properties, we can simplify the expression for the $I_{l,n}$ as follows.

PROPOSITION 4.7.

$$I_{l,n}(\theta, \phi) = \begin{cases} \sum_{m=0}^{+l} \frac{2}{1+\delta_{m,0}} N_{l,m} b_{l,n,m} P_{l,m}(\cos\theta) \cos m\phi, & l \text{ even}, \\ \sum_{m=1}^{+l} 2 N_{l,m} i b_{l,n,m} P_{l,m}(\cos\theta) \sin m\phi, & l \text{ odd}. \end{cases} \tag{9}$$

Proposition 4.7 implies that the ISBFs are completely determined by the $b_{l,n,m}$ coefficients for which $m \geq 0$. Therefore in the remainder of the paper, we assume $m \geq 0$ and $m' \geq 0$. Also, we absorb the "i" that occurs for l odd into $b_{l,n,m}$ so that the new definition of $b_{l,n,m}$ is always real (Proposition 4.6(2)). The calculation of the $b_{l,n,m}$ coefficients is the same in plan but different in details for l even versus l odd. We will show the l even case and then state the results for l odd.

Notice in equation (8) that the $b_{l,n,m}$ coefficients enter only through the quantity $\sum_{n=0}^{N_l-1} b_{l,n,m} b_{l,n,m'}$. Therefore, define

$$C_{l,m,m'} = \sum_{n=0}^{N_l-1} b_{l,n,m} b_{l,n,m'}. \tag{10}$$

The remainder of the calculation is in two parts: (1) explicit computation of the $C_{l,m,m'}$ constants and (2) factorization of (10) to determine the $b_{l,n,m}$ coefficients.

5. Calculation of $C_{l,m,m'}$. Denote the integer part function by $\lfloor\cdot\rfloor$. Use Proposition 4.7 in Proposition 4.5 and specialize to the case of l even, $m = 5\mu$ with $\mu = 0, \ldots, \lfloor l/5 \rfloor$ and $m' = 5\mu'$ with $\mu' = 0, \ldots, \lfloor l/5 \rfloor$ to get the result that

$$\begin{aligned}&\sum_{m'=0}^{l} C_{l,m,m'} \frac{2N_{l,m'}}{1+\delta_{m',0}} P_{l,m'}(\cos\theta_0)\cos m'\phi_0 \\ &= \frac{1}{6} N_{l,m}\left[P_{l,m}(\cos\theta_0)\cos m\phi_0 + \sum_{k=0}^{4} P_{l,m}(\cos\gamma_k)\cos m\alpha_k\right]. \end{aligned} \tag{11}$$

Multiply both sides of (11) by $\cos m''\phi_0$, integrate from 0 to 2π with respect to ϕ_0, and then divide by 2π. After using the orthonormality of $\cos m\phi_0$ and renaming m'' to m' we obtain Proposition 5.1, which is the basis for computing the $C_{l,m,m'}$.

PROPOSITION 5.1. *For l even, $m = 5\mu$ with $\mu = 0, \ldots, \lfloor l/5 \rfloor$ and $m' = 5\mu'$ with $\mu' = 0, \ldots, \lfloor l/5 \rfloor$,*

$$\begin{aligned}N_{l,m'} C_{l,m,m'} P_{l,m'}(\cos\theta_0) &= \frac{1}{6} N_{l,m}\left[P_{l,m'}(\cos\theta_0)\delta_{m,m'}\frac{1+\delta_{m',0}}{2}\right. \\ &\left. + \frac{1}{2\pi}\int_0^{2\pi}\sum_{k=0}^{4} P_{l,m}(\cos\gamma_k)\cos(m\alpha_k)\cos(m'\phi_0)\mathrm{d}\phi_0\right]. \end{aligned} \tag{12}$$

Equation (12) is of the form $C_{l,m,m'} f_{l,m'}(\theta_0) = h_{l,m,m'}(\theta_0)$. Therefore, for fixed l, m, and m', the functions $f_{l,m'}(\cdot)$ and $h_{l,m,m'}(\cdot)$ are proportional and $C_{l,m,m'}$ is the constant of proportionality. We are unable to compute the value of the integral contained in $h_{l,m,m'}(\cdot)$. However, we can compute $\lim_{\theta_0\to 0}(1/m'!)\mathrm{d}^{m'}/\mathrm{d}\theta_0^{m'}$ of both $f_{l,m'}(\cdot)$ and $h_{l,m,m'}(\cdot)$ and the resulting functions continue to have the same constant of proportionality. Define constants $g_{l,m'}$ and $D_{l,m,m'}$ by

$$g_{l,m'} = \left[\frac{1}{m'!}\frac{\mathrm{d}^{m'}}{\mathrm{d}\theta_0^{m'}} P_{l,m'}(\cos\theta_0)\right]_{\theta_0=0}, \tag{13}$$

$$D_{l,m,m'} = \sum_{k=0}^{4}\left[\frac{1}{m'!}\frac{\mathrm{d}^{m'}}{\mathrm{d}\theta_0^{m'}}\frac{1}{2\pi}\int_0^{2\pi} P_{l,m}(\cos\gamma_k)\cos(m\alpha_k)\cos(m'\phi_0)\mathrm{d}\phi_0\right]_{\theta_0=0}. \tag{14}$$

Substitute these definitions into the limit of the m'th derivative of (12) to obtain the final equation for determining $C_{l,m,m'}$ in terms of $g_{l,m'}$, $D_{l,m,m'}$, and the standard formula for $N_{l,m}$.

PROPOSITION 5.2. *For l even, $m = 5\mu$ with $\mu = 0, \ldots, \lfloor l/5 \rfloor$ and $m' = 5\mu'$ with $\mu' = 0, \ldots, \lfloor l/5 \rfloor$,*

$$N_{l,m'} C_{l,m,m'} g_{l,m'} = \frac{1}{6} N_{l,m}\left[g_{l,m'}\delta_{m,m'}\frac{1+\delta_{m',0}}{2} + D_{l,m,m'}\right]. \tag{15}$$

We are unable to directly evaluate the derivatives in (13) and (14) and then set $\theta_0 = 0$ so instead we use the following proposition.

PROPOSITION 5.3. *Let $f(\cdot)$ be a function with continuous arbitrary order derivatives. If $\lim_{\theta\to 0} f(\theta)/\theta^m = C$ and $|C| < \infty$, then*

$$\left[\frac{1}{m!}\frac{\mathrm{d}^m}{\mathrm{d}\theta^m} f(\theta)\right]_{\theta=0} = \lim_{\theta\to 0}\frac{f(\theta)}{\theta^m} = C. \tag{16}$$

```
Initialization: c_{l,m,-1} = 0, c_{m-1,m,m'} = 0.
for( m' = 0 ; m' ≤ M' ; m' ++ ){
        Compute z_{m'} using equation (40);
        for( m = m' ; m ≤ L ; m ++ ){
              Compute c_{m,m,m'} using equation (44);
              for( l = m + 1 ; l ≤ L ; l ++ ){
                    Compute c_{l,m,m'} using equation (42);
              }
        }
}
```

FIG. 1. *Recursive algorithm to compute* $c_{l,m,m'}$.

Using Proposition 5.3 we can evaluate $g_{l,m'}$ and $D_{l,m,m'}$ (see Appendices A and B) with the results that

$$g_{l,m'} = \frac{(-1)^{m'}(l+m')!}{2^{m'}m'!(l-m')!},$$
$$D_{l,m,m'} = \frac{5}{2^{m'}}c_{l,m,m'},$$

where $c_{l,m,m'}$ can be computed either explicitly by

$$c_{l,m,m'} = 2^{-l}\sum_{i=\lfloor\frac{l+m}{2}\rfloor}^{l}(-1)^{l-i}\frac{(2i)!}{(l-i)!i!}\left(\frac{1}{\sqrt{5}}\right)^{2i-l-m}\sum_{j=0}^{2i-l-m}\frac{2^j}{(2i-l-m-j)!j!}$$
$$\times\sum_{p=0,2,\ldots}^{m}\frac{m!}{p!}\left(\frac{2}{\sqrt{5}}\right)^{m-p}\sum_{q=0}^{m-p}\left(-\frac{1}{2}\right)^{q}\frac{\delta_{m',j+p+q}}{(m-p-q)!q!}$$

or recursively as shown in Figure 1.

6. Factorization of $C_{l,m,m'}$ to compute $b_{l,n,m}$. Once we have $C_{l,m,m'}$, we use (10) to calculate $b_{l,n,m}$ using well-known matrix factorization algorithms. Note that there is no interaction between different values of l in (10) and so in this section l takes some fixed value and that value is suppressed in the matrix notation. Let $\mathbf{C}$ and $\mathbf{b}$ be matrices of dimensions $\lfloor l/5\rfloor \times \lfloor l/5\rfloor$ and $N_l \times \lfloor l/5\rfloor$, respectively, in which the (n,μ)th elements are $C_{l,5n,5\mu}$ and $b_{l,n,5\mu}$, respectively. Equation (10) is then equivalent to

$$\mathbf{C} = \mathbf{b}^{\mathrm{T}}\mathbf{b}. \tag{17}$$

Therefore, $\mathbf{C}$ is symmetric and positive semidefinite. By orthonormality of ISBFs within the same l it follows that $\mathbf{b}\mathbf{b}^{\mathrm{T}} = \mathbf{I}_{N_l}$ (Proposition 4.2), where $\mathbf{I}_q$ is the $q \times q$ identity matrix and therefore $\mathbf{C}$ is also idempotent. Because $\mathbf{C}$ is idempotent, any factorization of $\mathbf{C}$ will be row orthonormal as described in Proposition 6.1.

PROPOSITION 6.1. *Let* $\mathbf{U} \in \mathcal{R}^{n\times n}$. *If* $\mathbf{V} \in \mathcal{R}^{m\times n}$ *is (row) full rank and* $\mathbf{U} = \mathbf{V}^{\mathrm{T}}\mathbf{V}$, *then* $\mathbf{V}$ *is row orthonormal if and only if* $\mathbf{U}$ *is idempotent.*

Note that if $\mathbf{b}$ is a solution to (17), then for any $N_l \times N_l$ orthogonal matrix $\mathbf{O}$, $\mathbf{b}' = \mathbf{O}\mathbf{b}$ is also a solution. For this reason we may add an additional constraint on $\mathbf{b}$ requiring it to be upper triangular, which implies $b_{l,n,5\mu} \equiv 0$ for $\mu < n$.

One algorithm to factor $\mathbf{C}$ is eigenvalue decomposition. Because $\mathbf{C}$ is idempotent $\mathbf{C}$ has only two eigenvalues, 0 and 1. Rows of $\mathbf{b}$ span the same space as eigenvectors of $\mathbf{C}$ with eigenvalues 1. Gram–Schmidt orthogonalization may be used on these eigenvectors to obtain orthogonal row vectors of $\mathbf{b}$. This algorithm requires all elements of $\mathbf{C}$ and it usually does not generate an upper triangular solution.

An alternative factorization algorithm is the Cholesky factorization, which is the algorithm we have used in computer codes. In this algorithm $b_{l,n,m}$ are computed by

$$b_{l,n,5n} = \sqrt{C_{l,5n,5n} - \sum_{n'=0}^{n-1} b_{l,n',5n}^2}, \tag{18}$$

$$b_{l,n,5n'} = \frac{1}{b_{l,n,5n}}\left(C_{l,5n',5n} - \sum_{k=0}^{n-1} b_{l,k,5n} b_{l,k,5n'}\right), \quad n' = n+1, \ldots, N_l - 1. \tag{19}$$

Equations (18) and (19) should be applied in the order $n = 0, 1, \ldots, N_l - 1$ to ensure that the $b_{l,m,m'}$ that occur on the right-hand side are already determined by the time they are needed. This algorithm requires only elements of $C_{l,m,m'}$ for the index values $0 \le m' \le 5N_l$ and $m' \le m \le l$. This is a computational advantage because computation of $C_{l,m,m'}$ can be expensive especially for large l, m, m'. The algorithm generates an upper triangular $\mathbf{b}$.

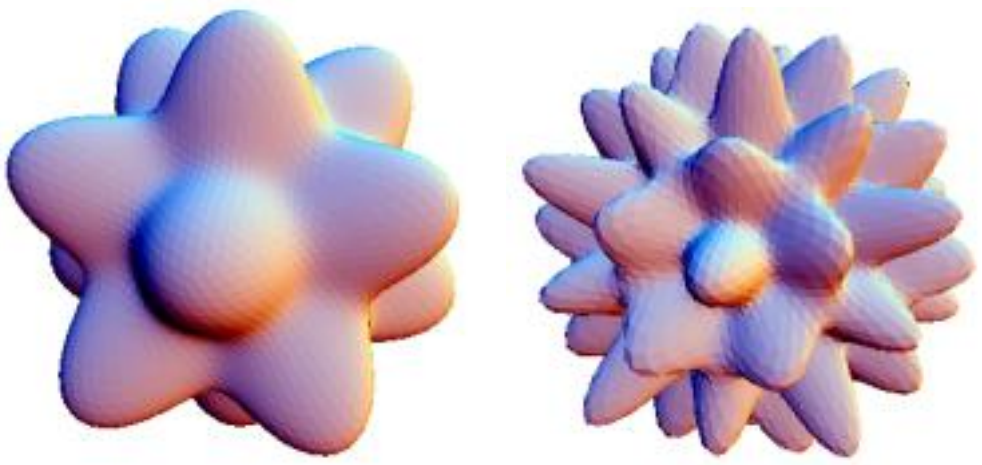
7. Numerical example. For $l \in \{0, 1, \ldots, 29\}$ there are either zero or one harmonic for each l and the cases with one harmonic are $l \in \{0, 6, 10, 12, 15, 16, 18, 20\text{–}22, 24\text{–}28\}$. By evaluating the recursions of this paper we have computed the harmonics through $l = 85$. Here we state only the first four harmonics in unnormalized form as computed symbolically by *Mathematica*:

$$\begin{aligned}
I_{0,0}(\theta,\phi) &= 1,\\
I_{6,0}(\theta,\phi) &= 2^3\cdot 3^2\cdot 5\cdot 11 P_{6,0}(\cos\theta) - P_{6,5}(\cos\theta)\cos 5\phi,\\
I_{10,0}(\theta,\phi) &= 2^8\cdot 3^4\cdot 5^2\cdot 7\cdot 13\cdot 19 P_{10,0}(\cos\theta) + 2^5\cdot 3^2\cdot 5\cdot 19 P_{10,5}(\cos\theta)\cos 5\phi\\
&\quad + P_{10,10}(\cos\theta)\cos 10\phi,\\
I_{12,0}(\theta,\phi) &= 2^8\cdot 3^5\cdot 5^2\cdot 7^2\cdot 11\cdot 17 P_{12,0}(\cos\theta) - 2^4\cdot 3^2\cdot 5\cdot 7\cdot 11 P_{12,5}(\cos\theta)\cos 5\phi\\
&\quad + P_{12,10}(\cos\theta)\cos 10\phi.
\end{aligned}$$

(Division of the stated formula by $\sqrt{2^2\pi}$, $2^4\cdot 3^2\cdot 5^2\sqrt{11\pi/13}$, $2^9\cdot 3^4\cdot 5^4\sqrt{7\cdot 13\cdot 19\pi}$, or $2^9\cdot 3^4\cdot 5^3\cdot 7\cdot 11\sqrt{5\cdot 7\cdot 17\pi}$ will normalize $I_{0,0}$, $I_{6,0}$, $I_{10,0}$, or $I_{12,0}$, respectively.) In Figure 2 we show a spherical plots of $I_{6,0}$ and $I_{12,0}$ which clearly exhibit the icosahedral symmetry of $I_{6,0}$ and $I_{12,0}$.

8. The l odd case. The calculations are similar to the case of l even. Here we list only the major results. The explicit expression for $c_{l,m,m'}$ (37) is modified to

$$\begin{aligned}
c_{l,m,m'} = (-1)2^{-l} &\sum_{i=\lfloor \frac{l+m}{2}\rfloor}^{l} (-1)^{l-i}\frac{(2i)!}{(l-i)!i!}\left(\frac{1}{\sqrt{5}}\right)^{2i-l-m}\sum_{j=0}^{2i-l-m}\frac{2^j}{(2i-l-m-j)!j!}\\
&\times \sum_{p=1,3,\ldots}^{m}\frac{m!}{p!}\left(\frac{2}{\sqrt{5}}\right)^{m-p}\sum_{q=0}^{m-p}\left(-\frac{1}{2}\right)^q\frac{\delta_{m',j+p+q}}{(m-p-q)!q!}.
\end{aligned}$$



$I_{6,0}(\theta,\phi)$ $I_{12,0}(\theta,\phi)$

FIG. 2. *The ISBFs for* $(l,n)=(6,0)$ *and* $(l,n)=(12,0)$. *For each value of* θ *and* ϕ *the distance of the surface from the origin is* $c_{l,n}+I_{l,n}(\theta,\phi)$, *where* $c_{l,n}=2\max_{\theta,\phi}(|I_{l,n}(\theta,\phi)|)$.

For the recursive calculations, the initial conditions of z_m are modified to $z_1=1, z_2=-1$ and (44) is modified to

$$c_{m,m,m'}=\left(-\frac{1}{\sqrt{5}}\right)^{m-1}\frac{(2m)!}{m!}\left(\frac{1}{2}\right)\binom{m}{m'}z_{m'}.$$

9. Generalization to other representations and other rotational groups. The idea of applying the projection operator to the delta function can be applied to other finite groups of coordinate rotations and to higher dimensional representations.

Let g be the order of the finite group $\mathcal{G}$ of coordinate rotations; N be the number of irreducible representations; and d_p, for $p=0,\ldots,N-1$, be the dimension of the pth irreducible representation. For the icosahedral group, these values are $g=60$, $N=5$, and $d_p=1,3,3,4,5$ [23, p. 324]. Let $\Gamma^p(T_k)_{j,j'}$ for $p=0,\ldots,N-1$, $k=0,\ldots,g-1$, and $j,j'=1,\ldots,d_p$ be the matrix elements of the kth member of the group in the pth unitary irreducible representation which, for the icosahedral group, are tabulated in [15].

We continue to use the notation and results of [22] specialized to square integrable functions on the sphere which are indicated by $L^2(\theta,\phi)$. Let $f(\theta,\phi)\in L^2(\theta,\phi)$. By [22, Theorem I, p. 92] it follows that

$$f(\theta,\phi)=\sum_{p=0}^{N-1}\sum_{j=0}^{d_p-1}a_j^p f_j^p(\theta,\phi), \tag{20}$$

where $f_j^p(\theta,\phi)$ is a normalized basis function transforming as the jth row of the d_p-dimensional unitary irreducible representation Γ^p of $\mathcal{G}$, a_j^p are a set of complex numbers, and the sum on p is over all the inequivalent unitary irreducible representations of $\mathcal{G}$. Following [22, p. 93] we define the projection operator $\mathcal{P}_{j,j'}^p$ by

$$\mathcal{P}_{j,j'}^p=\frac{d_p}{g}\sum_{T\in\mathcal{G}}\Gamma^p(T)_{j,j'}^* P(T).$$

By [22, Theorem II, p. 93] it follows that

$$\mathcal{P}^p_{j,j} f(\theta,\phi) = a^p_j f^p_j(\theta,\phi),$$

where a^p_j and $f^p_j(\theta,\phi)$ are the coefficients and basis functions of the expansion of $f(\theta,\phi)$ (20) that relate to the jth row of Γ^p.

We apply these results to $\delta(\cos\theta - \cos\theta_0)\delta(\phi-\phi_0)$ to find that

$$\begin{aligned}
\delta(\cos\theta-\cos\theta_0)\delta(\phi-\phi_0) &= \sum_{p=0}^{N-1}\sum_{j=0}^{d_p-1} a^p_j \Delta^p_j(\theta_0,\phi_0;\theta,\phi),\\
a^p_j\Delta^p_j(\theta_0,\phi_0;\theta,\phi) &= \mathcal{P}^p_{j,j}\delta(\cos\theta-\cos\theta_0)\delta(\phi-\phi_0)\\
&= \frac{d_p}{g}\sum_{k=1}^{g}\Gamma^p(T_k)^*_{j,j}\delta(\cos\theta-\cos\theta_k)\delta(\phi-\phi_k), \qquad (21)
\end{aligned}$$

where (θ_k,ϕ_k) are the symmetry-related positions, e.g., for the icosahedral group, (θ_k,ϕ_k) are given by Proposition 4.3. The normalization a^p_j is set by the condition $\int \Delta^p_j(\theta_0,\phi_0;\theta,\phi)\mathrm{d}\Omega = 1$.

The symmetrized delta functions $\Delta^p_j(\theta_0,\phi_0;\theta,\phi)$ define subspaces, denoted by $(L^p_j)^2(\theta,\phi)$, of the Hilbert space $L^2(\theta,\phi)$ by

$$(L^p_j)^2(\theta,\phi) = \left\{ f(\theta,\phi)\in L^2(\theta,\phi) : f(\theta,\phi) = \int \Delta^p_j(\theta_0,\phi_0;\theta,\phi) f(\theta_0,\phi_0)\mathrm{d}\Omega_0 \right\}.$$

Each subspace contains only a certain type of basis function, the union of the subspaces is all of $L^2(\theta,\phi)$, and the only function in the intersection of any pair of the subspaces is the zero function.

The goal is to determine a complete orthonormal fixed basis in each subspace. Denote the fixed basis functions by $I^p_j(\theta,\phi;\alpha)$ where α is an index. We proceed exactly as in the previous sections of the paper devoted to the identity representation of the icosahedral group. First, one can show that α can be written as l,n and

$$I^p_j(\theta,\phi;l,n) = \sum_{m=-l}^{+l} b^p_j(l,n,m) Y_{l,m}(\theta,\phi).$$

Second, one can expand $\Delta^p_j(\theta_0,\phi_0;\theta,\phi)$ as a weighted sum of $Y_{l,m}(\theta,\phi)$:

$$\Delta^p_j(\theta_0,\phi_0;\theta,\phi) = \sum_{l=0}^{\infty}\sum_{m=-l}^{+l} w^p_j(\theta_0,\phi_0;l,m) Y_{l,m}(\theta,\phi),$$

specifically (by using (21)),

$$\Delta^p_j(\theta_0,\phi_0;\theta,\phi) = \frac{d_p}{g a^p_j}\sum_{k=1}^{g}\Gamma^p(T_k)^*_{j,j}\sum_{l=0}^{\infty}\sum_{m=-l}^{+l} Y^*_{l,m}(\theta_k,\phi_k) Y_{l,m}(\theta,\phi). \qquad (22)$$

Third, since $I^p_j(\theta,\phi;l,n)$ are a complete orthonormal fixed basis for $(L^p_j)^2(\theta,\phi)$, it follows that

$$\Delta^p_j(\theta_0,\phi_0;\theta,\phi) = \sum_{l=0}^{\infty}\sum_{n=0}^{N^p_j(l)} \left(I^p_j(\theta_0,\phi_0;l,n)\right)^* I^p_j(\theta,\phi;l,n). \qquad (23)$$

Fourth, by equating the expansions for $\Delta_j^p(\theta_0, \phi_0; \theta, \phi)$ provided by (22) and (23), one arrives at an equation that is exactly a generalization of (5). From this point forward, the $I_j^p(\theta, \phi; l, n)$ can be obtained by using the same methods already used for the identity representation of the icosahedral group.

10. Conclusion. We have described a novel method for explicitly computing orthonormal symmetrized harmonics and have applied the method to the identity representation of the icosahedral group. The work was motivated by the analysis of data from spherical viruses. Other applications of icosahedral symmetry include fullerenes [24] and quasi crystals [10]. A *Mathematica* program to obtain exact closed-form expressions for ISBFs of arbitrary order and a C program to calculate their numerical values are available from the authors upon request.

The same approach can be used to determine general explicit expressions for other groups and the particular examples of tetrahedrally and octahedrally symmetric basis functions have been done by the authors. Moreover, since the icosahedrally symmetric delta function can be viewed as the result of applying the projection operator of the identity representation of the icosahedral group to the regular delta function, we believe the same technique can be employed to calculate basis functions for the other four irreducible representations of the icosahedral group. These functions are of great interest in quantum mechanical problems with an icosahedrally symmetric potential, of which one example is the C_{60} molecule. Work in this direction is already in progress and will be reported in a future publication.

Appendix A. Computation of $g_{l,m'}$. Since

$$P_{l,m}(x) = \frac{(-1)^m(1-x^2)^{m/2}}{2^l l!}\frac{\mathrm{d}^{l+m}}{\mathrm{d}x^{l+m}}(x^2-1)^l = (-1)^m(1-x^2)^{m/2}G_{l,m}(x), \tag{24}$$

where

$$G_{l,m}(x) = \frac{1}{2^l l!}\frac{\mathrm{d}^{l+m}}{\mathrm{d}x^{l+m}}(x^2-1)^l \tag{25}$$

is a polynomial of order $l-m$, we find that

$$P_{l,m'}(\cos\theta_0) = (-1)^{m'}(\sin\theta_0)^{m'}G_{l,m'}(\cos\theta_0).$$

Using Proposition 5.3 we obtain

$$g_{l,m'} = \lim_{\theta_0\to 0}\frac{P_{l,m'}(\cos\theta_0)}{\theta_0^{m'}} = (-1)^{m'}G_{l,m'}(1) = \frac{(-1)^{m'}(l+m')!}{2^{m'}m'!(l-m')!}.$$

Appendix B. Computation of $D_{l,m,m'}$. We begin the calculation of $D_{l,m,m'}$ by recalling the trigonometric and polynomial definitions of the Chebyshev polynomials of the first kind:

$$T_m(x) = \cos(m\arccos x) = \sum_{p=0,2,\ldots}^{m}\binom{m}{p}x^{m-p}(x^2-1)^{p/2}. \tag{26}$$

Define $R_{l,m}(x,y)$ by

$$\begin{aligned} R_{l,m}(x,y) &= P_{l,m}\big((y+2x)/\sqrt{5}\,\big)\cos\Big(m\arccos\Big((2-x)/\sqrt{5-(y+2x)^2}\Big)\Big) \\ &= \Big(-\frac{1}{\sqrt{5}}\Big)^m G_{l,m}\Big(\frac{1}{\sqrt{5}}(y+2x)\Big) \\ &\quad\times \sum_{p=0,2,\ldots}^{m}\binom{m}{p}(2-x)^{m-p}[5x^2+y^2-1+4x(y-1)]^{p/2}, \end{aligned} \tag{27}$$

where we have used (25) and (26). $R_{l,m}(x,y)$ derives its importance from the fact that $R_{l,m}(\sin\theta_0\cos\phi_k,\cos\theta_0) = P_{l,m}(\cos\gamma_k)\cos(m\alpha_k)$ which is central in the definition of $D_{l,m,m'}$ (14). Since $R_{l,m}(x,y)$ is a polynomial of order l in x and y it can be written in the form

$$R_{l,m}(x,y) = \sum_{m''=0}^{l} c_{l,m,m''}(y)x^{m''}, \tag{28}$$

where $c_{l,m,m''}(y)$ is a polynomial in y of order at most l.

PROPOSITION B.1. *Define* $A_{m'',m'}$ *by*

$$A_{m'',m'} = \frac{1}{2\pi}\int_0^{2\pi}(\cos\phi_k)^{m''}\cos(m'\phi_0)\mathrm{d}\phi_0. \tag{29}$$

Then

1. *if* $m'' < m'$, *then* $A_{m'',m'} = 0$,
2. $A_{m',m'} = (1/2^{m'})\cos\frac{2\pi}{5}km'$.

Define $Q_{l,m,m'}(\theta_0)$ by

$$Q_{l,m,m'}(\theta_0) = \frac{1}{2\pi}\int_0^{2\pi} R_{l,m}(\sin\theta_0\cos\phi_k,\cos\theta_0)\cos(m'\phi_0)\mathrm{d}\phi_0,$$

which is the first step on the path from $R_{l,m}(\sin\theta_0\cos\phi_k,\cos\theta_0)$ to $D_{l,m,m'}$. Note that all dependence of $R_{l,m}(\sin\theta_0\cos\phi_k,\cos\theta_0)$ on ϕ_k (and thus on ϕ_0) comes from the first argument x. Using Proposition B.1 and (28) we obtain

$$Q_{l,m,m'}(\theta_0) = \sum_{m''=m'}^{l} c_{l,m,m''}(\cos\theta_0)(\sin\theta_0)^{m''}A_{m'',m'}. \tag{30}$$

Furthermore, by Proposition 5.3 and (30),

$$\Big[\frac{1}{m'!}\frac{\mathrm{d}^{m'}}{\mathrm{d}\theta_0^{m'}}Q_{l,m,m'}(\theta_0)\Big]_{\theta_0=0} = \lim_{\theta_0\to 0}\frac{Q_{l,m,m'}(\theta_0)}{\theta_0^{m'}} = c_{l,m,m'}(1)\frac{1}{2^{m'}}\cos\frac{2\pi}{5}km'. \tag{31}$$

In addition, if m' is an integer multiple of 5, then $\sum_{k=0}^{4}\cos\frac{2\pi}{5}km' = 5$. Using this fact and (31) in the definition of $D_{l,m,m'}$ (14) we obtain

$$D_{l,m,m'} = \frac{5}{2^{m'}}c_{l,m,m'}(1). \tag{32}$$

It remains only to calculate $c_{l,m,m'}(1)$. In the following two subsections we provide two methods, a finite summation and a recurrence, both based on the observation that

$$R_{l,m}(x,1) = \sum_{m'=0}^{l} c_{l,m,m'}(1)x^{m'}. \tag{33}$$

That is, $c_{l,m,m'}(1)$ is the coefficient of the term $x^{m'}$ in the lth order polynomial $R_{l,m}(x,1)$. For notational convenience, from now on we shall rewrite $c_{l,m,m'}(1)$ as $c_{l,m,m'}$ and $R_{l,m}(x,1)$ as $R_{l,m}(x)$.

B.1. Explicit expression for $c_{l,m,m'}$. Substituting $y = 1$ into (27) we obtain

$$R_{l,m}(x) = \left(-\frac{2}{\sqrt{5}}\right)^m G_{l,m}\left(\frac{1}{\sqrt{5}}(1+2x)\right)H_m(x), \tag{34}$$

where

(35)

$$H_m(x) = (1-x-x^2)^{m/2}T_m\left(\frac{2-x}{2\sqrt{1-x-x^2}}\right) = \sum_{p=0,2,\ldots}^{m}\binom{m}{p}\left(1-\frac{x}{2}\right)^{m-p}\left(\frac{\sqrt{5}}{2}x\right)^p.$$

The function $G_{l,m}(\cdot)$ can be evaluated for an arbitrary argument from its definition in (25): take the derivative term by term of the binomial expansion of $(x^2-1)^l$ to obtain

$$G_{l,m}(x) = \frac{1}{2^l l!}\sum_{i=\lfloor\frac{l+m}{2}\rfloor}^{l}\binom{l}{i}(-1)^{l-i}\frac{(2i)!}{(2i-l-m)!}x^{2i-l-m}.$$

By further use of the binomial expansion, we can obtain the following expression for $R_{l,m}(x)$:

$$\begin{aligned} R_{l,m}(x) = 2^{-l}&\sum_{i=\lfloor\frac{l+m}{2}\rfloor}^{l}(-1)^{l-i}\frac{(2i)!}{(l-i)!i!}\left(\frac{1}{\sqrt{5}}\right)^{2i-l-m}\sum_{j=0}^{2i-l-m}\frac{2^j x^j}{(2i-l-m-j)!j!} \\ &\times\sum_{p=0,2,\ldots}^{m}\frac{m!}{p!}\left(\frac{2}{\sqrt{5}}\right)^{m-p}x^p\sum_{q=0}^{m-p}\left(-\frac{1}{2}\right)^q\frac{x^q}{(m-p-q)!q!}. \end{aligned} \tag{36}$$

From (36) and (33) it is clear that an explicit expression for $c_{l,m,m'}$ is

$$\begin{aligned} c_{l,m,m'} = 2^{-l}&\sum_{i=\lfloor\frac{l+m}{2}\rfloor}^{l}(-1)^{l-i}\frac{(2i)!}{(l-i)!i!}\left(\frac{1}{\sqrt{5}}\right)^{2i-l-m}\sum_{j=0}^{2i-l-m}\frac{2^j}{(2i-l-m-j)!j!} \\ &\times\sum_{p=0,2,\ldots}^{m}\frac{m!}{p!}\left(\frac{2}{\sqrt{5}}\right)^{m-p}\sum_{q=0}^{m-p}\left(-\frac{1}{2}\right)^q\frac{\delta_{m',j+p+q}}{(m-p-q)!q!}. \end{aligned} \tag{37}$$

B.2. Recursive calculation of $c_{l,m,m'}$. Using the recursive relation for Chebyshev polynomials

$$T_{m+1}(x) - 2xT_m(x) + T_{m-1}(x) = 0$$

and (35) we can derive the following recursive relation for $H_m(x)$:

$$H_{m+1}(x) + (x-2)H_m(x) + (1-x-x^2)H_{m-1}(x) = 0 \tag{38}$$

with the initial condition $H_0(x) = 1, H_1(x) = 1 - x/2$. The solution of (38) is

$$\begin{aligned} H_m(x) &= \frac{1}{2}\left[\left(1 + \frac{-1+\sqrt{5}}{2}x\right)^m + \left(1 - \frac{\sqrt{5}+1}{2}x\right)^m\right] \\ &= \frac{1}{2}\sum_{m'=0}^{m}\binom{m}{m'} z_{m'}x^{m'}, \end{aligned} \tag{39}$$

where

$$z_{m'} = \left(\frac{-1+\sqrt{5}}{2}\right)^{m'} + (-1)^{m'}\left(\frac{\sqrt{5}+1}{2}\right)^{m'}.$$

Note that $z_{m'}$ satisfies the recursion

$$z_{m'+1} + z_{m'} - z_{m'-1} = 0 \tag{40}$$

with the initial condition $z_0 = 2$ and $z_1 = -1$.

Using (34), (24), and the recursion for $P_{l,m}(x)$ in l

$$(l+1-m)P_{l+1,m}(x) - (2l+1)xP_{l,m}(x) + (l+m)P_{l-1,m}(x) = 0,$$

we can derive a recursion for $R_{l,m}(x)$ in l:

$$(l+1-m)R_{l+1,m}(x) - (2l+1)\frac{1}{\sqrt{5}}(1+2x)R_{l,m}(x) + (l+m)R_{l-1,m}(x) = 0. \tag{41}$$

Substitute (33) into (41). The coefficient of each power of x must vanish separately, which leads to the recursion

$$(l+1-m)c_{l+1,m,m'} - (2l+1)\frac{1}{\sqrt{5}}(c_{l,m,m'} + 2c_{l,m,m'-1}) + (l+m)c_{l-1,m,m'} = 0. \tag{42}$$

To initialize (42) to compute $c_{l,m,m'}$, note that $c_{l,m,-1} = 0$ and $c_{m-1,m,m'} = 0$. We still need $c_{m,m,m'}$ to start the recursion, but

$$\sum_{m'=0}^{m} c_{m,m,m'}x^{m'} = R_{m,m}(x) = \left(-\frac{1}{\sqrt{5}}\right)^m \frac{(2m)!}{m!}H_m(x) \tag{43}$$

so by (39)

$$c_{m,m,m'} = \left(-\frac{1}{\sqrt{5}}\right)^m \frac{(2m)!}{m!}\left(\frac{1}{2}\right)\binom{m}{m'}z_{m'}. \tag{44}$$

Given an integer L, Figure 1 shows an algorithm, with control structures written in the C programming language, to calculate all $c_{l,m,m'}$ for $l \le L, m' \le M' = 5N_L, m' \le m \le l$. Because the factorization algorithm described in section 6 uses only $C_{l,m,m'}$ for which $0 \le m' \le 5N_l, m' \le m \le l$, it follows that for any single l we need only to compute $c_{l,m,m'}$ for $0 \le m' \le 5N_l, m' \le m \le l$.

Acknowledgments. The authors are grateful to an anonymous reviewer for pointing out [19]. The authors were motivated by discussions with Professor John E. Johnson (Department of Molecular Biology, The Scripps Research Institute) regarding the modeling of spherical viruses.

REFERENCES

[1] W. Chiu, R.M. Burnett, and R.L. Garcea, eds., *Structural Biology of Viruses*, Oxford University Press, London, 1997.

[2] T.S. Baker and J.E. Johnson, *Low resolution meets high: Towards a resolution continuum from cells to atoms*, Curr. Opin. Struc. Biol., 6 (1996), pp. 585–594.

[3] Y. Zheng and P.C. Doerschuk, *Iterative reconstruction of three-dimensional objects from averaged Fourier-transform magnitude: Solution and fiber x-ray scattering problems*, J. Opt. Soc. Amer. A, 13 (1996), pp. 1483–1494.

[4] Y. Zheng and P.C. Doerschuk, *3D image reconstruction from averaged Fourier transform magnitude by parameter estimation*, IEEE Trans. Image Process., 7 (1998), pp. 1561–1570.

[5] Y. Zheng, P.C. Doerschuk, and J.E. Johnson, *Determination of three-dimensional low-resolution viral structure from solution x-ray scattering data*, Biophys. J., 69 (1995), pp. 619–639.

[6] J.D. Jackson, *Classical Electrodynamics*, John Wiley, New York, 2nd ed., 1975.

[7] S.L. Altmann, *On the symmetries of spherical harmonics*, Proc. Cambridge Philos. Soc., 53 (1957), pp. 343–367.

[8] B. Meyer, *On the symmetries of spherical harmonics*, Canad. J. Math., 135 (1954), pp. 135–157.

[9] N.V. Cohan, *The spherical harmonics with the symmetry of the icosahedral group*, Proc. Cambridge Philos. Soc., 54 (1958), pp. 28–38.

[10] L. Elcoro, J.M. Perez-Mato, and G. Madariaga, *Determination of quasicrystalline structures: A refinement program using symmetry-adapted parameters*, Acta Cryst. Sect. A, 50 (1994), pp. 182–193.

[11] E. Heuser-Hofmann and W. Weyrich, *Three-dimensional reciprocal form factors and momentum densities of electrons from Compton experiments:* I. *Symmetry-adapted series expansion of the electron momentum density*, Z. Naturforsch., 40a (1985), pp. 99–111.

[12] A. Jack and S.C. Harrison, *On the interpretation of small-angle x-ray solution scattering from spherical viruses*, J. Mol. Biol., 99 (1975), pp. 15–25.

[13] M. Kara and K. Kurki-Suonio, *Symmetrized multipole analysis of orientational distributions*, Acta Cryst. Sect. A, 37 (1981), pp. 201–210.

[14] O. Laporte, *Polyhedral harmonics*, Z. Naturforsch., 3a (1948), pp. 447–456.

[15] F. Liu, J.-L. Ping, and J.-Q. Chen, *Application of the eigenfunction method to the icosahedral group*, J. Math. Phys., 31 (1990), pp. 1065–1075.

[16] A.G. McLellan, *Eigenfunctions for integer and half-odd integer values of J symmetrized according to the icosahedral group and the group C_{3v}*, J. Chem. Phys., 34 (1961), pp. 1350–1359.

[17] J. Raynal, *Determination of point group harmonics for arbitrary j by a projection method.* II. *Icosahedral group, quantization along an axis of order* 5, J. Math. Phys., 25 (1984), pp. 1187–1194.

[18] Y. Zheng and P.C. Doerschuk, *Explicit orthonormal fixed bases for spaces of functions that are totally symmetric under the rotational symmetries of a Platonic solid*, Acta Cryst. Sect. A, 52 (1996), pp. 221–235.

[19] G. W. Fernando, M. Weinert, R.E. Watson, and J.W. Davenport, *Point group symmetries and Gaussian integration*, J. Comput. Phys., 112 (1994), pp. 282–290.

[20] J.T. Finch and K.C. Holmes, *Structural studies of viruses*, in Methods in Virology, Vol. III, Academic Press, New York, London, 1967, chapter 9.

[21] M.E. Rose, *Elementary Theory of Angular Momentum*, John Wiley, New York, 1957.

[22] J.F. Cornwell, *Group Theory in Physics*, Vol. 1, Academic Press, New York, London, 1984.

[23] M. Artin, *Algebra*, Prentice-Hall, Englewood Cliffs, NJ, 1991.

[24] W. Kratschmer, L.D. Lamb, K. Fostiropoulos, and D.R. Huffman, *Solid C_{60}: A new form of carbon*, Nature (London), 347 (1990), pp. 354–357.

SIAM J. MATH. ANAL.
Vol. 32, No. 3, pp. 555–571

PERIODIC SOLUTIONS OF HAMILTONIAN SYSTEMS*

YANHENG DING† AND CHENG LEE‡

Abstract. Two sequences of periodic solutions with large and small norms, respectively, are obtained for Hamiltonian systems of the type

$$-\mathcal{J}\dot{z} = \xi F_z(t,z) + \eta G_z(t,z),$$

where F is superquadratic at $z=\infty$ and G is subquadratic at $z=0$.

Key words. Hamiltonian system, periodic solution, variational method

AMS subject classifications. 34B30, 34C25

PII. S0036141099358178

1. Introduction and statement of results. In recent years, several papers Ambrosetti, Brézis, and Cerami [2], Ambrosetti, Azorero, and Peral [1], and Bartsch and Willem [4] investigated the structure of solutions of semilinear elliptic equations of the type

$$Au = \xi\nabla\Phi(u) + \eta\nabla\Psi(u),$$

where A is the Laplacian on a smoothly bounded domain $\Omega \subset \mathbb{R}^N$, $\xi, \eta \in \mathbb{R}$ are given constants, $\Phi \in \mathcal{C}^1(H_0^1(\Omega), \mathbb{R})$ is superquadratic at $u=\infty$, and $\Psi \in \mathcal{C}^1(H_0^1(\Omega), \mathbb{R})$ is subquadratic at $u=0$. The results have also been extended to problems of periodic solutions of some special Hamiltonian systems in [4] and of wave equations, etc., in [6, 7].

Hamiltonian systems have been extensively studied via critical point theory since the publication of the pioneer work by Rabinowitz [15] (see [5] and references therein). In the present paper we consider the existence of infinitely many periodic solutions for the following Hamiltonian system:

$$\dot{z} = \mathcal{J}\nabla_z H_{\xi\eta}(t,z), \tag{HS$_{\xi\eta}$}$$

where $\mathcal{J} := \left(\begin{smallmatrix} 0 & -I \\ I & 0 \end{smallmatrix}\right)$ denotes the standard symplectic matrix on $\mathbb{R}^{2N}$, and $H_{\xi\eta}$ is of the form

$$H_{\xi\eta}(t,z) = \xi F(t,z) + \eta G(t,z), \tag{1.1}$$

where ξ, η are real constants and with F, $G \in \mathcal{C}^1(\mathbb{R}\times\mathbb{R}^{2N})$ being T-periodic in t.

To establish the existence of solutions with large norms, we mainly need the following assumptions on F:

*Received by the editors June 18, 1999; accepted for publication (in revised form) April 18, 2000; published electronically September 15, 2000.

http://www.siam.org/journals/sima/32-3/35817.html

†Institute of Mathematics, Academia Sinica, Beijing, China (dingyh@math03.math.ac.cn). The work of this author was supported by the Alexander von Humboldt-Stiftung and partly by NSFC 19971091.

‡Department of Mathematics, National Changhua University of Education, Changhua, Taiwan, China (clee@math.ncue.edu.tw). The work of this author was supported by the National Science Council, Taiwan (NSC89-2115-M-018-011).

(F1) there are $\mu > 2$, $\bar{r} > 0$ such that

$$0 < \mu F(t,z) \le F_z(t,z)z \quad \text{for all } |z| \ge \bar{r};$$

(F2) there are $\nu > 2$ and a_1, $a_2 > 0$ such that, letting $\nu' := \nu/(\nu-1)$,

$$|F_z(t,z)|^{\nu'} \le a_1 + a_2 F_z(t,z)z \quad \text{for all } (t,z);$$

and on G

(G1) there are $\tau \in [0, 1)$ and a_3, $a_4 > 0$ such that

$$|G(t,z)| + |G_z(t,z)z| \le a_3 + a_4|F(t,z)|^\tau \quad \text{for all } (t,z).$$

For obtaining solutions with small energies we assume

(F3) $F(t,0) = 0$ and $F_z(t,z) = o(|z|)$ uniformly in t as $z \to 0$;

(G2) $G(t,0) = 0$, $G_z(t,z)z/|z|^2 \to \infty$ as $z \to 0$ uniformly in t, and there are $1 < \alpha < 2 < \beta$ and $\underline{r}$, $a_5 > 0$ such that

$$G_z(t,z)z \le \alpha G(t,z) \quad \text{for all } t \in \mathbb{R} \text{ and } 0 < |z| \le \underline{r},$$
$$a_5|G_z(t,z)|^\beta \le G(t,z) \quad \text{for all } (t,z) \in \mathbb{R} \times B_{\underline{r}},$$

where $B_{\underline{r}} := \{z \in \mathbb{R}^{2N} : \ |z| \le \underline{r}\}$.

Letting $|u|_p$ denote the usual $L^p([0,T],\mathbb{R}^{2N})$-norm for $p \in [1, \infty]$ and $I_{\xi\eta}(u)$ the energy for a T-periodic solution u of $(\mathrm{HS}_{\xi\eta})$,

$$I_{\xi\eta}(u) = \int_0^T \left(\frac{1}{2}\dot{u}\mathcal{J}u - H_{\xi\eta}(t,u)\right) dt,$$

and the first result reads as follows.

THEOREM 1.1. *Let $H_{\xi\eta}(t,z)$ be of the form* (1.1).

(a) *Assume* (F1), (F2),

(F4) $F(t,-z) = F(t,z)$ *for all* (t,z),

and (G1) *with* $\tau < 2/\nu$. *Then for every* $\xi \ne 0$, $\eta \in \mathbb{R}$, $(\mathrm{HS}_{\xi\eta})$ *has a sequence of* T*-periodic solutions* (u_n) *satisfying* $\xi I_{\xi\eta}(u_n) \to \infty$ *and* $|u_n|_\infty \to \infty$ *as* $n \to \infty$.

(b) *Assume* (F3), (F4), (G2) *and*

(G3) $G(t,-z) = G(t,z)$ *for all* $(t,z) \in \mathbb{R} \times B_{\underline{r}}$.

Then for every $\xi \in \mathbb{R}$, $\eta \ne 0$, $(\mathrm{HS}_{\xi\eta})$ *has a sequence of* T*-periodic solutions* (v_n) *satisfying* $\eta I_{\xi\eta}(v_n) < 0, \eta I_{\xi\eta}(v_n) \to 0$, *and* $|v_n|_\infty \to 0$ *as* $n \to \infty$.

(c) *If, moreover,* $H_{\xi\eta}$ *can be represented as* $H_{\xi\eta} = a(t)\hat{H}_{\xi\eta}(z)$ *with* $\hat{H}_{\xi\eta}(z) \ne 0$ *almost everywhere (a.e.) in* z *and* $a(t)$ *having minimal period* T, *then each solution of the sequences* (u_n) *and* (v_n) *has also minimal period* T.

The next result is related to the situation where F is independent of t.

THEOREM 1.2. *Let $H_{\xi\eta}(t,z)$ be of the form* (1.1).

(a) *Assume* (F1), (F2),

(F5) $F(t,z) = F(z)$, *i.e., F is independent of t,*

and (G1) *with* $\tau < 2/\nu$. *Then the conclusion of Theorem* 1.1(a) *holds. If, in addition,* $\eta \ne 0$, $G(t,z) = a(t)\hat{G}(z)$ *with* $\hat{G}(z) \ne 0$ *a.e. in* z *and* $a(t)$ *having minimal period* T, *then each* u_n *has also minimal period* T.

(b) *Assume* (F3), (F5), (G2), *and*

(G4) $G(t,z) = G(z)$ *for all* $z \in B_{\underline{r}}$.

Then, for all $T > 0$, $(HS_{\xi\eta})$ has a sequence of T-periodic solutions (v_n) satisfying $0 < -\eta I_{\xi\eta}(v_n) \to 0$ and $0 < \inf_{t\in[0,T]} |v_n(t)| \le |v_n|_\infty \to 0$ as $n \to \infty$.

Remarks. Some remarks are in order.

(1) First, part (a) of Theorems 1.1 and 1.2 are results of symmetries with perturbation. In this direction, it seems that there are no papers discussing the case where F does depend on t and is even in z. If F is independent of t, many results are known; see, e.g., [3, 15, 12, 13], etc., where G is bounded jointly sometimes with certain restrictions on $\partial G/\partial t$. In particular, a result of [13] applies to the case where $F(z)$ satisfies (F1), and there are $2 < p_1 \le p_2 \le 2p_1$, $1 \le p < 2p_1/p_2$, $q > 0$ such that $a_1|z|^{p_1} - a_2 \le F(z) \le a_3|z|^{p_2} + a_4$, $|G(t,z)| \le a_5(|z|^p + 1)$, $|G_z(t,z)| \le a_5(|z|^{q-1} + 1)$, and $|G_t(t,z)| \le a_5(|z|^q + 1)$. Clearly this does not cover, e.g., the case $F(z) = |z|^\mu + |z|^{3\mu}$ and $G(t,z) \sim |z|^{1/3}$ near $z = \infty$, which is indeed contained in our results.

(2) In part (b) of Theorems 1.1 and 1.2 the Hamiltonian $H_{\xi\eta}$ may not be defined on whole $\mathbb{R}^{2N}$, i.e., it is sufficient for obtaining the results if $H_{\xi\eta}$ is defined locally in a neighborhood of $0 \in \mathbb{R}^{2N}$ in which the symmetry and subquadratic conditions are satisfied. When it is given on $\mathbb{R}^{2N}$, we may ignore its states (symmetric or not, growing slowly or fast as $|z| \to \infty$) outside an arbitrarily small neighborhood of $0 \in \mathbb{R}^{2N}$. More precisely, our proof in section 5 together with [10] shows that the system

$$\dot z = \mathcal{J} G_z(t,z)$$

has a sequence of T-periodic solutions (v_n) satisfying $0 > \int_0^T \left(\frac{1}{2}\dot v_n \mathcal{J} v_n - G(t,v_n)\right) dt \to 0$ provided that

(i) there exist $\alpha \in (1,2)$, $\beta \ge \alpha$, and $\underline{r}$, $c > 0$ such that $G \in \mathcal{C}^1(\mathbb{R}\times B_{\underline{r}}, \mathbb{R})$, T-periodic in t, $G(t,0) = 0$, and, whenever $z \ne 0$,

$$0 < G_z(t,z)z \le \alpha G(t,z) \quad \text{and} \quad c\,|G_z(t,z)|^\beta \le G(t,z);$$

(ii) either $G(t,z)$ is even in $z \in B_{\underline{r}}$ or $G(t,z) = G(z)$ is independent of t.

(3) In part (b) of Theorem 1.2 the solution sequence (v_n) can be chosen so that the corresponding minimal period sequence (T_n) (where $T_n > 0$ denotes the minimal period of v_n) satisfying $T_n \to 0$. Indeed, given $T > 0$, let v_1 be a T-periodic solution with minimal period $T_1 = T/k_1$ for some $k_1 \in \mathbb{N}$ and $0 < -\eta I_{\xi\eta}(v_1), |v_1|_\infty \le 1$. Then let v_2 be a $T_1/2$-periodic solution with minimal period $T_2 = T_1/2k_2 = T/2k_1k_2$ for some $k_2 \in \mathbb{N}$ and $0 < -\eta I_{\xi\eta}(v_2), |v_2|_\infty \le 1/2$. Inductively, we can choose a $T_{n-1}/2$-periodic solution v_n with minimal period $T_n = T_{n-1}/2k_n = T/2^{n-1}k_1\cdots k_n$ for some $k_n \in \mathbb{N}$ and $0 < -\eta I_{\xi\eta}(v_n), |v_n|_\infty \le 1/n$. Clearly, v_n is T-periodic and $T_n \to 0, 0 < -\eta I_{\xi\eta}(v_n) \to 0, |v_n|_\infty \to 0$ as $n \to \infty$.

(4) It seems interesting to make a comparison with a result of Rabinowitz [16] where he proved that if $H(z) := H_{\xi\eta}(z)$ is superquadratic at $z = \infty$ (i.e., $0 < \mu H(z) \le \nabla H(z)z$ for all $|z|$ large), then for all $T > 0, R > 0$, $(\mathrm{HS}_{\xi\eta})$ possesses a T-periodic solution Z_T with $|z_T|_\infty \ge R$. Theorem 1.2(b), on the other hand, says that if $H(z)$ is subqudratic at $z = 0$, then for all $T > 0, \varepsilon > 0$, $(\mathrm{HS}_{\xi\eta})$ has a T-periodic solution z_T with $|z_T|_\infty \le \varepsilon$.

(5) Clearly, it may happen that the Hamiltonian $H_{\xi\eta}$ satisfies both the superquadratic (at $z = \infty$) and the subquadratic (at $z = 0$) conditions so that the system $(\mathrm{HS}_{\xi\eta})$ has the two sequences of periodic solutions. For this case, compared to the result of [4], in part (a) of Theorems 1.1 and 1.2 the subquadratic term G is no longer symmetric.

(6) Part (c) of Theorem 1.1 is a consequence of the following proposition. (The second conclusion of Theorem 1.2(a) can be verified similarly.)

PROPOSITION 1.3. *Suppose* $H \in \mathcal{C}^1(\mathbb{R}^{2N}, \mathbb{R})$, $H_z(z) \neq 0$ *for* $z \neq 0$, *and* $a \in \mathcal{C}(\mathbb{R}, \mathbb{R})$ *has the minimal period* T. *If* u *is a nonconstant* T*-periodic solution of*

$$-\mathcal{J}\dot{z} = a(t)H_z(z), \tag{1.2}$$

then u *has the minimal period* T.

Proof. See [10]. By the existence and uniqueness theorem of ordinary differential equations, $u(t) \neq$ any equilibrium everywhere. Let T_0 denote the minimal period of u. Using (1.2) one obtains $a(t+T_0)H_z(u(t)) = a(t)H_z(u(t))$ and so $a(t+T_0) = a(t)$ for all t, which clearly implies that $T_0 = T$. □

(7) Theorem 1.1(b) was established essentially in [10]. Thus it remains to prove part (a) of Theorems 1.1, 1.2 and part (b) of Theorem 1.2.

Apart from the preliminaries in section 2, we prove part (a) of Theorems 1.1, 1.2 in sections 3 and 4 by using Rabinowitz variational setting. Finally in section 5, of special interest is that a new index theory is developed, which is similar to Benci's pseudoindex. We use this new index theory, together with some approximation arguments, to prove part (b) of the theorems.

2. Preliminaries. Without loss of generality, assume $T = 2\pi$ for notational convenience. Let $S^1 := \mathbb{R}/[0, 2\pi]$, $L^2 := L^2(S^1, \mathbb{R}^{2N})$ with the usual inner product $(\cdot,\cdot)_{L^2}$ and norm $|\cdot|_2$, $W^{1,2} := W^{1,2}(S^1, \mathbb{R}^{2N})$, and consider the self-adjoint operator $A := -\mathcal{J}d/dt$ acting on L^2 with domain $\mathcal{D}(A) = W^{1,2}$. The spectrum $\sigma(A) = \mathbb{Z}$ and each $k \in \mathbb{Z}$ is an eigenvalue of multiplicity $2N$ corresponding to the eigenfunction $e^{kt\mathcal{J}}e_j$ for $j = 1, \dots, 2N$, where $e^{kt\mathcal{J}} := \cos kt\, I + \sin kt\, \mathcal{J}$ and $e_1, \dots, e_{2N}$ denote the usual orthogonal basis of $\mathbb{R}^{2N}$. Let $E(k)$ denote the eigenspace associated to eigenvalue k. Then

$$\begin{aligned} E(k) &= e^{kt\mathcal{J}}\mathbb{R}^{2N} = \left\{e^{kt\mathcal{J}}c :\ c \in \mathbb{R}^{2N}\right\} \\ &= \left\{\sum_{j=1}^{2N} c_j e^{kt\mathcal{J}}e_j :\ c_j \in \mathbb{R} \text{ for all } j = 1, \dots, 2N\right\}. \end{aligned}$$

Each $u \in L^2$ has the expression

$$u = \sum_{k\in\mathbb{Z}} e^{kt\mathcal{J}}c_k(u) = \sum_{j\in\mathbb{Z}}\sum_{j=1}^{2N} c_{kj}(u)e^{kt\mathcal{J}}e_j, \qquad c_k(u) \in \mathbb{R}^{2N},\ c_{kj}(u) \in \mathbb{R}.$$

We have the orthogonal decomposition

$$L^2 = L^- \oplus L^0 \oplus L^+, \qquad u = u^- + u^0 + u^+,$$

where $L^0 = E(0) = \mathbb{R}^{2N}$, and $L^\pm$ is the closure of $\oplus_{k\in\mathbb{N}}E(\pm k)$ in L^2. Let $E := \mathcal{D}(|A|^{1/2}) = W^{1/2,2}(S^1, \mathbb{R}^{2N})$ equipped with the inner product

$$\begin{aligned} (u, v) &:= 2\pi c_0(u)c_0(v) + 2\pi\sum_{k\in\mathbb{Z}} |k|c_k(u)c_k(v) \\ &= 2\pi\sum_{j=1}^{2N} c_{0j}(u)c_{0j}(v) + 2\pi\sum_{k\in\mathbb{Z}}\sum_{j=1}^{2N} |k|c_{kj}(u)c_{kj}(v) \end{aligned}$$

and norm $\|u\|^2 = (u,u)$. There holds the decomposition

$$E = E^- \oplus E^0 \oplus E^+ \quad \text{with} \quad E^0 = E(0) \text{ and } E^\pm = E \cap L^\pm,$$

orthogonal with respect to both $(\cdot,\cdot)_{L^2}$ and $(\cdot,\cdot)$. It is known that E is compactly embedded in L^p for all $p \in [1,\infty)$.

LEMMA 2.1. *For each $p \in [2,\infty)$ there is $C_p > 0$ such that*

$$|u|_p \le C_p m^{-1/p}\|u\|$$

for all $u \in (\oplus_{k=-m+1}^{m-1} E(k))^\perp$, the orthogonal complement in E, where (and below) $|\cdot|_p$ denotes the usual L^p-norm.

Proof. If $p = 2$, the conclusion is clear. Suppose $p > 2$. For $u = \sum_{|j|\ge m} e^{jt\mathcal{J}} c_j(u) \in E$, by the Hausdorff–Young and Hölder inequalities,

$$\begin{aligned} |u|_p &\le c_p \left(\sum_{|j|\ge m} |c_j|^q \right)^{1/q} \\ &\le c_p' \left(\sum_{|j|\ge m} |j|^{-q/(2-q)} \right)^{(2-q)/2q} \|u\|, \end{aligned}$$

where $1/p + 1/q = 1$. Since

$$\sum_{|j|\ge m} |j|^{-q/(2-q)} \le 2\int_m^\infty x^{-q/(2-q)}dx = \frac{2-q}{q-1} m^{-2(q-1)/(2-q)},$$

the lemma follows. □

Next, for later use in estimating a certain index we consider the following operator on L^2,

$$A_{V,\vartheta} u := V(t) \sum_{k\in\mathbb{Z}} \vartheta_k e^{kt\mathcal{J}} c_k(u) = V(t) \sum_{k\in\mathbb{Z}} \sum_{j=1}^{2N} \vartheta_k c_{kj}(u) e^{kt\mathcal{J}} e_j,$$

where $V : S^1 \to \mathbb{R}$, a given real function, and $\vartheta = (\vartheta_k) \in \ell^p(\mathbb{Z})$, the Banach space of real number sequences with $|\vartheta|_p = \left(\sum_{k\in\mathbb{Z}} |\vartheta_j|^p\right)^{1/p} < \infty$ $(p \ge 1)$. Note that both the multiplication operator V with $\mathcal{D}(V) = \{u \in L^2 : Vu \in L^2\}$ and the operator ϑ defined by $\vartheta e^{kt\mathcal{J}} e_j = \vartheta_k e^{kt\mathcal{J}} e_j$ are self-adjoint. Moreover, $\vartheta : L^2 \to L^2$ is compact and it is not difficult to check that if $V \in L^p, \vartheta \in \ell^p$ $(p \ge 2)$, then $\vartheta(L^2) \subset \mathcal{D}(V), A_{V,\vartheta} = V \circ \vartheta, A_{V,\vartheta}^* = \vartheta \circ V$, and $A_{V,\vartheta} : L^2 \to L^2$ is compact with $\|A_{V,\vartheta}\| \le C_p |V|_p |\vartheta|_p$. Let $l_1(V,\vartheta) \ge l_2(V,\vartheta) \ge \cdots$ denote the repeated eigenvalues of $|A_{V,\vartheta}| := (A_{V,\vartheta}^* A_{V,\vartheta})^{1/2}$. Assuming $V \in L^p$ and $\vartheta \in \ell^p$, recall that the "p-norm" is defined by

$$\|A_{V,\vartheta}\|_p := \left(\sum_n l_n(V,\vartheta)^p \right)^{1/p} \quad \text{for } p \in [1,\infty) \quad \text{and} \quad \|A_{V,\vartheta}\|_\infty := \|A_{V,\vartheta}\|.$$

It is known that

$$\|A_{V,\vartheta}\|_p^2 = \|A_{V,\vartheta}^* A_{V,\vartheta}\|_{p/2} \quad \text{for } p \ge 2 \tag{2.1}$$

and for any complete orthogonal sequence (ψ_n)

$$\|A_{V,\vartheta}\|_2 = \left(\sum_n |A_{V,\vartheta}\psi_n|_2^2\right)^{1/2} \tag{2.2}$$

as well as, letting $\mathcal{B}$ denote the family of orthogonal sequences in L^2,

$$\|A_{V,\vartheta}\|_p = \sup_{(\phi_n),(\psi_n)\in\mathcal{B}} \Big(|(\phi_n, A_{V,\vartheta}\psi_n)_{L^2}|^p\Big)^{1/p}. \tag{2.3}$$

Let $\{\lambda_n(V,\vartheta)\}$ be the set of all eigenvalues of $A^*_{V,\vartheta}A_{V,\vartheta}$, and $\mathcal{N}(V,\vartheta)$ the number of eigenvalues of $A^*_{V,\vartheta}A_{V,\vartheta}$ which are ≥ 1. We will apply the following.

LEMMA 2.2 (cf. [18]). *For $p \in [2,\infty], V \in L^p$, and $\vartheta \in \ell^p$, there is $C_p > 0$ such that*

$$\mathcal{N}(V,\vartheta) \leq C_p |V|_p^p |\vartheta|_p^p.$$

Proof. Using (2.2) with the basis $(e^{kt\mathcal{J}}e_j)$ one gets easily that

$$\|A_{V,\vartheta}\|_2 = \sqrt{2N}\,|V|_2|\vartheta|_2,$$

and plainly there holds

$$\|A_{V,\vartheta}\|_\infty \leq |V|_\infty |\vartheta|_\infty.$$

Now fixing arbitrarily $(\psi_n),(\phi_n) \in \mathcal{B}$, the operator $(V,\vartheta) \to ((\phi_n, A_{V,\vartheta}\psi_n))$ from $L^p \times \ell^p$ into ℓ^p satisfies

$$|(\phi_n, A_{V,\vartheta}\psi)|_2 \leq \|A_{V,\vartheta}\|_2 \leq \sqrt{2N}\,|V|_2|\vartheta|_2,$$
$$|(\phi_n, A_{V,\vartheta}\psi_n)|_\infty \leq \|A_{V,\vartheta}\|_\infty \leq |V|_\infty|\vartheta|_\infty.$$

Thus by a complex interpolation one sees for $p \in (2,\infty)$

$$|(\phi_n, A_{V,\vartheta}\psi_n)|_p \leq C_p|V|_p|\vartheta|_p$$

with C_p independent of $(\phi_n),(\psi_n) \in \mathcal{B}$, which, together with (2.3), implies

$$\|A_{V,\vartheta}\|_p \leq C_p|V|_p|\vartheta|_p \quad \text{for all } p \in [2,\infty]. \tag{2.4}$$

(2.1) and (2.4) then yield

$$(\mathcal{N}(V,\vartheta))^{2/p} \leq \left(\sum_n \lambda_n(V,\vartheta)^{p/2}\right)^{2/p} = \|A^*_{V,\vartheta}A_{V,\vartheta}\|_{p/2} = \|A_{V,\vartheta}\|_p^2 \leq C_p^2|V|_p^2|\vartheta|_p^2,$$

and so the desired result follows. □

Finally, recall that the group $\mathcal{G} = \mathbb{Z}/2 =: \{\mathrm{id}, -\mathrm{id}\}$ acts on L^2 by

$$T_{id}\,u = u \qquad \text{and} \qquad T_{-id}\,u = -u \tag{2.5}$$

and the group $\mathcal{G} = S^1 := \{e^{i\theta} :\ \theta \in [0, 2\pi)\}$, where $i = \sqrt{-1}$, acts on L^2 by

$$(T_\theta\,u)(t) = u(t+\theta), \tag{2.6}$$

so that L^2 becomes a $\mathcal{G}$-space. Since these actions commute with A, the eigenspace $E(k)$ is invariant under each of the actions for all k.

When L^2 is regarded as a $\mathbb{Z}/2$-space, we arrange the positive eigenvalues of A by (repeated in multiplicity) $1 \le \lambda_1 \le \lambda_2 \le \cdots$ corresponding to the eigenfunctions ω_n (i.e., $A\omega_n = \lambda_n \omega_n$) and set

$$E_k = E(k) \text{ for } k \le 0 \text{ and } E_k = \mathbb{R}\,\omega_k \text{ for } k \in \mathbb{N}. \tag{2.7}$$

Then each $u \in L^2$ has the expression $u = u^- + u^0 + \sum_{k\in\mathbb{N}} c_k(u)\omega_k$ with $c_k(u) \in \mathbb{R}$.

When L^2 is regarded as a S^1-space, it is convenient to consider the complexification $\mathbb{R}^{2N} \cong \mathbb{C}^N$ and $L^2 \cong L^2(S^1, \mathbb{C}^N)$, and to set

$$E_k = e^{ikt}\mathbb{C}^N \text{ for } k \le 0 \text{ and } E_k = e^{i\bar{k}t}\mathbb{C} \text{ for } k \in \mathbb{N}, \text{ where } \bar{k} := \left[\frac{k+N-1}{N}\right]. \tag{2.8}$$

Note that, in this case, each $u \in L^2$ has the expression

$$u(t) = \sum_{k\le 0} e^{ikt}c_k(u) + \sum_{k\ge 1} e^{i\bar{k}t}c_k(u),\ c_k(u) \in \mathbb{C}^N$$

(regarding $\mathbb{C}$ as a subspace of $\mathbb{C}^N$), and the S^1-action reads as

$$(T_\theta\, u)(t) = \sum_{k\le 0} e^{ik\theta}e^{ikt}c_k(u) + \sum_{k\ge 1} e^{i\bar{k}\theta}e^{i\bar{k}t}c_k(u).$$

We will also consider another action on L^2 defined by

$$(\hat{T}_\theta\, u)(t) := c_0(u) + \sum_{k<0} e^{i\theta}e^{ikt}c_k(u) + \sum_{k>0} e^{i\theta}e^{i\bar{k}t}c_k(u). \tag{2.9}$$

Throughout the paper we denote

$$E_l^m := \oplus_{k=l}^m E_k, \quad E_{-l} := E_{-l}^0 \oplus E^+, \quad \text{and} \quad E^m := E^- \oplus E^0 \oplus E_1^m,$$

where it is understood that E_k takes different forms in different situations corresponding to (2.7) and (2.8), respectively.

3. Rabinowitz variational setting. In this section we always assume (F1)–(F2) and (G1) are satisfied. Observe that (F1)–(F2) imply that

$$c_1|z|^\mu \le F(t,z) \le c_2|z|^\nu \quad \text{whenever } |z| \ge \bar{r}, \tag{3.1}$$

where (and below) c_j's denote positive constants. Set

$$\Phi(u) := \int_0^{2\pi} F(t,u) \text{ and } \Psi(u) := \int_0^{2\pi} G(t,u).$$

Then $\Phi,\ \Psi \in \mathcal{C}^1(E,\mathbb{R})$ and

$$\Phi'(u)u \ge \mu\Phi(u) - c_3 \quad \text{for all } u \in E, \tag{3.2}$$

$$|\Psi(u)| + |\Psi'(u)u| \le c_3 + c_4|\Phi(u)|^\tau \quad \text{for all } u \in E. \tag{3.3}$$

Consider the functional

$$I_{\xi\eta}(u) := \frac{1}{2}(\|u^+\|^2 - \|u^-\|^2) - \xi\Phi(u) - \eta\Psi(u).$$

Then $I_{\xi\eta} \in \mathcal{C}^1(E, \mathbb{R})$ and its critical points give rise to periodic solutions of $(\mathrm{HS}_{\xi\eta})$.

We will consider only the case where $\xi > 0$ and $\eta \in \mathbb{R}$. The other case can be dealt with by considering the functional $-I_{\xi\eta}$ similarly. Furthermore, noting that (F1)–(F2) and (G1) remain true if F, G are replaced by ξF, $|\eta| G(t, z)$, respectively, with the constants a_j $(j = 1, 2, 3, 4)$ depending possibly on ξ, η, we can assume, for notational simplification and without loss of generality, that $\xi = 1$, $\eta = 1$, and denote simply $I(u) = I_{\xi\eta}(u)$.

It is easy to check using (3.1)–(3.3) that there exist $\alpha_1 > 0$, $\Lambda > 0$ such that if $u \in E$ satisfies

$$I_0(u) := \frac{1}{2}(\|u^+\|^2 - \|u^-\|^2) - \Phi(u) \geq \frac{\mu + 2}{4\mu}\Phi'(u)u - \Phi(u) - |\Psi'(u)u|,$$

then

$$1 \leq \Phi(u) + \alpha_1 \leq \Lambda(I_0(u)^2 + 1)^{1/2}. \tag{3.4}$$

Note in particular that if $u \in \mathcal{K}(I) := \{u \in E : I'(u) = 0\}$, then (3.4) holds. Indeed, since

$$I(u) = I(u) - \frac{1}{2}I'(u)u = \frac{1}{2}\Phi'(u)u - \Phi(u) + \frac{1}{2}\Psi'(u)u - \Psi(u),$$

we have

$$I_0(u) = \frac{1}{2}\Phi'(u)u - \Phi(u) + \frac{1}{2}\Psi(u) - \Psi(u).$$

Set for $u \in E$

$$Q(u) := 2\Lambda(I_0(u)^2 + 1)^{1/2}, \quad \theta(u) := Q(u)^{-1}(\Phi(u) + \alpha_1), \quad \text{and} \quad \rho(u) := \chi(\theta(u)),$$

where $\chi \in \mathcal{C}^\infty(\mathbb{R}, \mathbb{R})$ such that $\chi(s) = 1$ for $s \leq 1, \chi(s) = 0$ for $s \geq 2$, and $\chi'(s) \in (-2, 0)$ for $s \in (1, 2)$. Consider the functional $J \in \mathcal{C}^1(E, \mathbb{R})$ defined by

$$J(u) := \frac{1}{2}(\|u^+\|^2 - \|u^-\|^2) - \Phi(u) - \rho(u)\Psi(u) = I_0(u) - \rho(u)\Psi(u).$$

By assumption, $|J(u) - J(T_g u)| = \rho(u)|\Psi(u) - \Psi(T_g u)|$ for all $g \in \mathcal{G}$, where $\mathcal{G}$ stands for the $\mathbb{Z}/2$ or S^1 according to (F4) or (F5), respectively, and, using (3.2)–(3.4), it is not difficult to check that

$$|J(u) - J(T_g u)| \leq \alpha_2(|J(u)|^\tau + 1) \ \text{ for all } u \in E \text{ and } g \in \mathcal{G}. \tag{3.5}$$

Moreover, one can verify the following (cf. [17, Proposition 10.16]).

LEMMA 3.1. *There exists $M > 0$ such that $J(u) = I(u)$ for u in a neighborhood of $\mathcal{K}(J) \cap J^{-1}[M, \infty)$.*

Therefore, it is sufficient for obtaining large norm solutions to show that J possesses an unbounded sequence of positive critical values.

In the following let $J_l := J|_{E_{-l}}$, the restriction of J on E_{-l}. Recall that a sequence $(u_j) \subset E$ (resp., E_{-l}) is called a $(PS)_c$-sequence for J (resp., J_l) if it satisfies $J(u_j) \to c$ and $J'(u_j) \to 0$ (resp., $J_l'(u_j) \to 0$), and a sequence (u_l) with $u_l \in E_{-l}$ is called a $(PS)_c^*$-sequence for J if it satisfies $J(u_l) \to c$ and $J_l'(u_l) \to 0$. J (resp., J_l) is said to satisfy $(PS)_c$ condition if any $(PS)_c$-sequence for J (resp., J_l)

has a convergent subsequence, and J is said to satisfy $(PS)^*_c$ condition if any $(PS)^*_c$-sequence for J has a convergent subsequence. We can verify easily by (3.1)–(3.3) the following lemma.

LEMMA 3.2. *There is a positive constant denoted again by M such that for all $c \geq M$, J and J_l satisfy $(PS)_c$ condition, and J also satisfies $(PS)^*_c$ condition.*

By (3.1)–(3.3) there are again positive constants γ_j such that for all $u \in E$

$$\frac{1}{2}\big(\|u^+\|^2 - \|u^-\|^2\big) - \gamma_1|u|_\nu^\nu - \gamma_2 \leq J(u) \leq \frac{1}{2}\big(\|u^+\|^2 - \|u^-\|^2\big) - \gamma_3|u|_\mu^\mu + \gamma_4. \tag{3.6}$$

Using this fact and Lemma 2.1, we have the following.

LEMMA 3.3. *For each $n \in \mathbb{N}$ there are $R_n > r_n > 0$ such that*

$$\sup J(E^n \backslash B_{R_n}) \leq 0 \quad \textit{and} \quad \alpha_n := \sup J(E^n) < \infty; \tag{3.7}$$

$$\inf J(\partial B_{r_n} \cap (E^{n-1})^\perp) \geq \gamma_5 n^{2/(\nu-2)}. \tag{3.8}$$

Without loss of generality we assume $R_n < R_{n+1}$ for all n. Let

$$D_l^n := B_{R_n} \cap E_{-l}^n,$$
$$\Gamma_l^n := \{\gamma \in \mathcal{C}(D_l^n, E_{-l}) : \ \gamma \text{ satisfies } (\gamma_1) \text{ and } (\gamma_2)\},$$

where

(γ_1) $\gamma = \mathrm{id}$ on $(\partial B_{R_n} \cap E_{-l}^n) \cup (D_l^n \cap \mathrm{Fix}\mathcal{G})$;

(γ_2) γ is $\mathcal{G}$-invariant in the following sense (cf. (2.5) and (2.6), (2.9)):

$$\begin{aligned} &\gamma(-u) = -\gamma(u) \ \text{ if } \mathcal{G} = \mathbb{Z}/2, \\ &\gamma(\hat{T}_\theta u) = T_\theta \gamma(u) \ \text{ for all } \theta \in [0,\, 2\pi) \ \text{ if } \mathcal{G} = S^1. \end{aligned} \tag{3.9}$$

Set also, taking $\phi_k \in E_k$ with $\|\phi_k\| = 1$ for all $k \in \mathbb{N}$,

$$U_l^n := \{u \in D_l^{n+1} : \ u = x + c\phi_{n+1} \text{ with } x \in E_{-l}^n \text{ and } c \geq 0\},$$
$$\Lambda_l^n := \{\lambda \in \mathcal{C}(U_l^n, E_{-l}) : \ \lambda|_{D_l^n} \in \Gamma_l^n, \lambda = id \text{ on } Q_l^n\},$$

where $Q_l^n := (U_l^n \cap \partial B_{R_{n+1}}) \cup ((U_l^n \cap E_{-l}^n) \backslash D_l^n) \cup (U_l^n \cap \mathrm{Fix}\mathcal{G})$. Define the minimax values for J_l:

$$b_l^n := \inf_{\gamma \in \Gamma_l^n} \sup_{u \in D_l^n} J(\gamma(u)),$$

$$c_l^n := \inf_{\lambda \in \Lambda_l^n} \sup_{u \in U_l^n} J(\lambda(u)).$$

LEMMA 3.4. *There holds*

$$\gamma_5 n^{2/(\nu-2)} \leq b_l^n \leq c_l^n \leq \alpha_{n+1} \quad \textit{for all } n, l \in \mathbb{N}.$$

Proof. Plainly, by definition and (3.7), $b_l^n \leq c_l^n \leq \alpha_n$. Set for $\gamma \in \Gamma_l^n$

$$\mathcal{O} := \{u \in D_l^n : \ \|\gamma(u)\| < r_n\}.$$

Since $0 \in \mathrm{Fix}\mathcal{G}$ and $\gamma = \mathrm{id}$ on $\partial D_l^n \cup (D_l^n \cap \mathrm{Fix}\mathcal{G})$, $\mathcal{O} \cap \partial D_l^n = \emptyset$, $\mathcal{O}$ is an open bounded invariant neighborhood of 0 in E_{-l}^n. Let $P : E_{-l}^n \to E_{-l}^{n-1}$ be the orthogonal

projection, and consider the composion $P\circ\gamma : \mathcal{O} \to E^{n-1}_{-l}$ which clearly keeps $\mathcal{O}\cap \mathrm{Fix}\mathcal{G}$ fixed. By the Borsuk–Ulam theorem (for its S^1-version, see [11]), $P\circ\gamma$ has a zero $u \in \partial\mathcal{O}$. Thus $\|\gamma(u)\| = r_n$ and $\gamma(u) \in (E^{n-1})^\perp$. Now (3.8) applies. □

By virtue of this lemma there exists a subsequence denoted again by $(l) \subset \mathbb{N}$ such that for all $n \in \mathbb{N}$, as $l \to \infty$,

$$b^n_l \to b_n, \quad c^n_l \to c_n, \quad \text{and} \quad \gamma_5 n^{2/(\nu-2)} \le b_n \le c_n \le \alpha_{n+1}. \tag{3.10}$$

LEMMA 3.5. *If $c_n > b_n \ge M$, then J has a critical value $\hat{c}_n \in [b_n, c_n]$.*

Proof. See [17] for the $\mathbb{Z}/2$-action case. The result for the S^1-action case can be proved similarly. Since $b_n < c_n$, we have $b^n_l < c^n_l$ for all l large. Let $\delta \in (0, c^n_l - b^n_l)$ and

$$\Lambda^n_l(\delta) := \{\lambda \in \Lambda^n_l : \ J(\lambda(u)) \le b^n_l + \delta \text{ for all } u \in D^n_l\},$$
$$c^n_l(\delta) := \inf_{\lambda\in\Lambda^n_l(\delta)} \sup_{u\in U^n_l} J(\lambda(u)).$$

Since J_l satisfies $(PS)_c$ for $c \ge M$, a standard deformation argument shows that $c^n_l(\delta)$ is a critical value of J_l. Then by $(PS)^*_c$ for $c \ge M$ one sees that J has a critical value in $[b_n, c_n]$. □

LEMMA 3.6. *If $b_n = c_n$ for all $n \ge \hat{n}$, then*

$$b_n \le \gamma_6 n^{1/(1-\tau)} \quad \textit{for all } n \in \mathbb{N}.$$

Proof. By the choice of E_k (cf. (2.7)–(2.8)) $D^{n+1}_l = \cup_{g\in\mathcal{G}} T_g(U^n_l)$ and for any $u \in D^{n+1}_l \backslash E^n_{-l}$ there is a unique $(x, g) \in (U^n_l \backslash D^n_l) \times \mathcal{G}$ such that $T_g x = u$. Each $\lambda \in \Lambda^n_l$ extends to $\gamma \in \Gamma^{n+1}_l$ via $\gamma(u) = T_g(\lambda(T^{-1}_g u))$, and thus by (3.5)

$$\begin{aligned} b^{n+1}_l &\le \sup_{u\in D^{n+1}_l} J(\gamma(u)) = \sup_{u\in U^n_l, g\in\mathcal{G}} J(T_g\lambda(u)), \\ &\le \sup_{u\in U^n_l} \big(J(\lambda(u)) + \alpha_2(|J(\lambda(u))|^{1/(1-\tau)} + 1)\big). \end{aligned}$$

Therefore for all n, l

$$b^{n+1}_l \le c^n_l + \alpha_2(|c^n_l|^{1/(1-\tau)} + 1)$$

which induces as $l \to \infty$

$$b_{n+1} \le b_n + \alpha_2(b_n^{1/(1-\tau)} + 1) \quad \text{for } n \ge \hat{n}.$$

Now the conclusion follows as done in [17]. □

4. Large norm solutions. We now turn to prove part (a) of the theorems.

Proof of Theorems 1.1(a) and 1.2(a). Observe that (3.6) implies on E^+

$$J(u) \ge f(u) - \gamma_2, \qquad \text{where} \qquad f(u) := \frac{1}{2}\|u\|^2 - \gamma_1|u|^\nu_\nu \in \mathcal{C}^2(E^+, \mathbb{R}).$$

Let $f_m := f|_{E^m_1}$, the restriction of f on E^m_1. Plainly, f and f_m satisfy the $(PS)_c$ and f satisfies the $(PS)^*_c$ for all c. Let BL denote the unit sphere of the vector space L and $SL := \partial(BL)$. We set for $n, m \in \mathbb{N}$ with $n < m$

$$A^m_n := \{\sigma \in \mathcal{C}(S\mathbb{R}^{m-n+1}, E^m_1) : \ \sigma(-x) = -\sigma(x) \text{ for all } x\} \tag{4.1}$$

if $\mathcal{G} = \mathbb{Z}/2$ is concerned (for proving Theorem 1.1(a)) and

$$A_n^m := \{\sigma \in \mathcal{C}(S\mathbb{C}^{m-n+1}, E_1^m) : \ \sigma(\hat{T}_\theta z) = T_\theta \sigma(z) \text{ for all } (z,\theta)\} \tag{4.2}$$

if $\mathcal{G} = S^1$ is concerned (for proving Theorem 1.2(a)). Define

$$\beta_n^m := \sup_{\sigma \in A_n^m} \min_{x \in SL} J(\sigma(x)),$$

where $L = \mathbb{R}^{m-n+1}$ and $L = \mathbb{C}^{m-n+1}$ corresponding to (4.1) and (4.2), respectively. Then it is easy to see that

$$\gamma_5 n^{2/(\nu-2)} \le \beta_n^m \le \alpha_n, \tag{4.3}$$

$$\beta_n^m \le \beta_{n+1}^m, \tag{4.4}$$

and β_n^m is a critical value of f_m. We can choose a sequence $m_j \to \infty$ such that the following limit exists:

$$\beta_n = \lim_{j\to\infty} \beta_n^{m_j} \qquad \text{for each } n \in \mathbb{N}.$$

Since f satisfies $(PS)_c^*$, each β_n is a critical value of f. Moreover, $\beta_n \le \beta_{n+1}$ by (4.4), and $\beta_n \to \infty$ by (4.3). We have the following claims:

1. $b_n \ge \beta_n - \gamma_2$ for all $n \in \mathbb{N}$;
2. along a subsequence $n_j \to \infty$ there is $u_{n_j} \in \mathcal{K}(f)$ satisfying

$$\beta_{n_j} \ge f(u_{n_j}) \ge C_p n_{n_j}^{p\nu/2(\nu-2)} \quad \text{for all } j,$$

where $C_p > 0$ independent of (u_{n_j}).

Postponing to show claims 1, 2, we first complete the proofs of Theorems 1.1(a) and 1.2(a). By the assumption $\tau < 2/\nu$, we have $1/(1-\tau) < p\nu/2(\nu-2)$ for $p > 2$ but closing sufficiently to 2. Then for j large enough, $b_{n_j} \ge \gamma_6 n_j^{p\nu/2(\nu-2)}$ in virtue of claims 1 and 2, which, jointly with Lemmas 3.4, 3.6, and 3.5 implies that J has an unbounded sequence of positive critical values, and thus we have part (a).

We now turn to the claims. Only the case of S^1-action is considered here because the other case can be dealt with similarly.

For proving claim 1 it is sufficient to show that for any $\gamma \in \Gamma_l^n$ and $\sigma \in A_n^m$, the intersection

$$\gamma(D_l^n) \cap \sigma(S\mathbb{C}^{m-n+1}) \ne \emptyset. \tag{4.5}$$

To verify this, let $P_l^m : E_{-l} \to E_{-l}^m$ be the orthogonal projection, and consider the composition $g := P_l^m \circ \gamma : D_l^n \to E_{-l}^m$. By a S^1-version of Borsuk–Ulam theorem (cf. [11] and [19]), $g(D_l^n) \cap \sigma(S\mathbb{C}^{m-n+1}) \ne \emptyset$. Thus there is $u_m \in D_l^n$ and $z_m \in S\mathbb{C}^{m-n+1}$ such that $P_l^m(\gamma(u_m)) = \sigma(z_m)$. Since D_l^n and $S\mathbb{C}^{m-n+1}$ are compact, along a subsequence as $m \to \infty, \gamma(u_m) \to \gamma(u), \sigma(z_m) \to \sigma(z)$ with $u \in D_l^n, z \in S\mathbb{C}^{m-n+1}$ and $\gamma(u) = \sigma(z)$. (4.5) is proved.

Next we prove claim 2. Recall that the index of a functional $g \in \mathcal{C}^2(X, \mathbb{R})$, X a Hilbert space, at $x \in X$ is defined by

$$\begin{aligned} \operatorname{Ind}(g,x) :&= \#\{\lambda : \lambda \le 0 \text{ is an eigenvalue of } g''(x)\} \\ &= \max\{\dim H : \ H \subset X, \text{ a subspace on which } g''(x) \text{ is nonpositive}\}. \end{aligned}$$

First we show that if $\beta_n < \beta_{n+1}$, then there is $u_n \in E^+$ satisfying

$$f(u_n) \le \beta_n, \quad f'(u_n) = 0, \quad \text{and} \quad \mathrm{Ind}(f, u_n) \ge 2n - 1. \tag{4.6}$$

Since $\beta_n < \beta_{n+1}$, we have $\beta_n^{m_j} < \beta_{n+1}^{m_j}$ for m_j large. By a result of Marino–Prodi (cf. [9, 18]), for any $0 < \varepsilon < (\beta_{n+1}^{m_j} - \beta_n^{m_j})/8$, we can take $\varphi_\varepsilon \in \mathcal{C}(E_1^{m_j}, \mathbb{R})$ which satisfies $(PS)_c$, has only finitely many nondegenerate critical points, and

$$\varphi_\varepsilon(u) = f(u) \text{ if } \|u - \mathcal{K}(f_{m_j})\| \ge \varepsilon, \qquad \|\varphi_\varepsilon - f_{m_j}\|_{\mathcal{C}^2} \le \varepsilon.$$

Define

$$\beta_n^{m_j}(\varepsilon) := \sup_{\sigma \in A_n^{m_j}} \min_{u \in S\mathbb{C}^{m-n+1}} \varphi_\varepsilon(\sigma(u)).$$

Take a regular value of φ_ε, $a_\varepsilon \in (\beta_n^{m_j} + 5\varepsilon,\, \beta_n^{m_j} + 6\varepsilon)$. Then

$$\beta_n^{m_j}(\varepsilon) < a_\varepsilon - 2\varepsilon < a_\varepsilon < \beta_{n+1}^{m_j}(\varepsilon),$$

and the homotopy group

$$\pi_{2(m_j-n)-1}([\varphi_\varepsilon \ge a_\varepsilon], p) \ne 0 \quad \text{for some } p \in [\varphi_\varepsilon \ge a_\varepsilon], \tag{4.7}$$

where $[\varphi_\varepsilon \ge a_\varepsilon] := \{u \in E_1^{m_j} :\ \varphi_\varepsilon(u) \ge a_\varepsilon\}$ (cf. [3]). On the other hand,

$$\pi_l([\varphi_\varepsilon \ge a_\varepsilon], p) = 0 \ \text{ for all } p \in [\varphi_\varepsilon \ge a_\varepsilon] \text{ and } l \le 2m_j - L(\varphi_\varepsilon, a_\varepsilon) - 1, \tag{4.8}$$

where $L(\varphi_\varepsilon, a_\varepsilon) := \max\{\mathrm{Ind}(\varphi_\varepsilon, u) :\ u \in \mathcal{K}(\varphi_\varepsilon), \varphi_\varepsilon(u) \le a_\varepsilon\}$. (4.7) and (4.8) imply that there is $u_\varepsilon \in \mathcal{K}(\varphi_\varepsilon)$ such that $\varphi_\varepsilon(u_\varepsilon) \le a_\varepsilon$ and $\mathrm{Ind}(\varphi_\varepsilon, u_\varepsilon) > 2n - 1$. Therefore, by $(PS)_c$, along a subsequence as $\varepsilon \to 0$, $u_\varepsilon \to u_n^{m_j} \in \mathcal{K}(f_{m_j})$ satisfying $f(u_n^{m_j}) \le \beta_n^{m_j}$ and $\mathrm{Ind}(f_{m_j}, u_n^{m_j}) \ge 2n - 1$. Now, by $(PS)_c^*$, letting $j \to \infty$ along a subsequence yields (4.6).

Let $u \in \mathcal{K}(f)$. Note that $(f''(u)v, v) = \|u\|^2 - \gamma_1\nu(\nu-1)(|u|^{\nu-2}v, v)_{L^2}$. Having this in mind we consider the operator $A_{\hat{V},\hat{\vartheta}}$ defined before (cf. section 2), where $\hat{\vartheta} = (\vartheta_k)$ with $\vartheta_k = 0$ for all $k \le 0$ and $\vartheta_k = k^{-1/2}$ for all $k \in \mathbb{N}$, and $\hat{V} = (\gamma_1\nu(\nu-1)|u|^{\nu-2})^{1/2}$. Clearly $\hat{V} \in L^p$ and $\hat{\vartheta} \in \ell^p$ for all $p > 2$, and the restriction of the operator $\hat{\vartheta}$ on L^+ is an isometry from L^+ to E^+. Letting $v := \hat{\vartheta}h$ for $h \in L^+, (f''(u)v, v) \le 0$ if and only if $(A_{\hat{V},\hat{\vartheta}}h, h)_{L^2} \ge |h|_2^2$. Therefore, $\mathrm{Ind}(f, u) = \mathcal{N}(\hat{V}, \hat{\vartheta})$, and so by Lemma 2.2,

$$\mathrm{Ind}(f, u) \le C_p|\hat{V}|_p^p|\hat{\vartheta}|_p^p \le C_p'||u|^{(\nu-2)/2}|_p^p \le C_p''|u|_\nu^{p(\nu-2)/2} \ \text{ for all } p > 2. \tag{4.9}$$

Since $f'(u) = 0$, one has $f(u) = f(u) - 1/2f'(u)u = (\nu/2 - 1)\gamma_1|u|_\nu^\nu$. In particular, we have by (4.6) and (4.9) (for $n \ge 2$)

$$\beta_n \ge f(u_n) = \frac{\nu - 2}{2}\gamma_1|u_n|_\nu^\nu \ge C_p n^{2\nu/p(\nu-2)},$$

that is, claim 2.

The proofs are completed. □

5. Solutions with small energies. We consider only the case where $\eta > 0$. (F3) and (G2) imply that there exist $\gamma \in [\alpha, 2)$, $r_0 \in (0, \underline{r}]$, and $a_6 > 0$ such that

$$0 < \nabla_z H_{\xi\eta}(t,z)z \le \gamma H_{\xi\eta}(t,z) \qquad \text{for all } t \in \mathbb{R} \text{ and } z \in B_{r_0}\backslash\{0\}, \tag{5.1}$$

$$a_6|\nabla_z H_{\xi\eta}(t,z)|^\beta \le H_{\xi\eta}(t,z) \qquad \text{for all } t \in \mathbb{R} \text{ and } z \in B_{r_0}. \tag{5.2}$$

Let $\chi = \chi(s) \in \mathcal{C}^\infty(\mathbb{R}, [0,1])$ be such that $\chi(s) = 0$ for $s \le r_0/2, \chi(s) = 1$ for $s \ge r_0$, and

$$\chi'(s) > 0 \qquad \text{for all } s \in (r_0/2, r_0). \tag{5.3}$$

Set $M = \inf\{H_{\xi\eta}(t,z)/r_0^\gamma : \ t \in \mathbb{R} \text{ and } |z| = r_0\}$. Consider $\tilde{H}_{\xi\eta} : \mathbb{R} \times \mathbb{R}^{2N} \to \mathbb{R}$ defined by

$$\tilde{H}_{\xi\eta}(t,z) = (1 - \chi(|z|))H_{\xi\eta}(t,z) + \chi(|z|)M|z|^\gamma.$$

Then by definition and (5.1)–(5.3),

$$\tilde{H}_{\xi\eta}(t,z) \ge M|z|^\gamma \qquad \text{for all } (t,z); \tag{5.4}$$

$$0 < \nabla_z \tilde{H}_{\xi\eta}(t,z)z \le \gamma \tilde{H}_{\xi\eta}(t,z) \qquad \text{for all } z \ne 0. \tag{5.5}$$

Without loss of generality, we assume $T = 2\pi$. Define $I \in \mathcal{C}^1(E, \mathbb{R})$ by

$$I(u) = \int_0^{2\pi} \tilde{H}_{\xi\eta}(t,u)dt - \frac{1}{2}\big(\|u^+\|^2 - \|u^-\|^2\big).$$

Then each critical point u of I with $|u|_\infty < r_0/2$ is a 2π-periodic solution of $(\mathrm{HS}_{\xi\eta})$.

We will give only the proof of Theorem 1.2(b). For the details of proof corresponding to the $\mathbb{Z}/2$-action case (i.e., Theorem 1.1(b)) we refer to [10]. Recall that in [10], a sequence $0 < c_n \to 0$ of minimax values for I was obtained as limits $c_n = \lim_{l\to\infty} c_l^n$, where $c_l^n := \sup\{\inf I(A \cap E_{-l}) : A \text{ is closed, symmetric, and } \mathbb{Z}/2\text{-genus}(A \cap E_{-l}) \ge 2N(l+n+1)\}$ (here $E_{-l} = \oplus_{j\ge -l}E(j)$). However, this does not work in the S^1-action situation. Indeed, for any $R > 0$, the S^1-genus$(S_R E(0)) = \infty$, and so the similar minimax values $= \infty$. Thus we have to try another way.

LEMMA 5.1. *For each $n \in \mathbb{N}$, there are $r_m > 0, \alpha_m > 0$ and $0 < \beta_m \to 0$ such that*

(i) $\inf I\big(S_{r_m\delta}E^m\big) \ge \alpha_m\delta^2 r_m^2$ *for all* $\delta \in [0,1]$;

(ii) $\sup I\big((E^{m-1})^\perp\big) \le \beta_m$.

Proof. Using the direct sum decomposition and the fact that $\dim(E^0 \oplus E_1^m) < \infty$, one has $|u|_\gamma^\gamma \ge c_1|u^-|_\gamma^\gamma + c_2\|u^0\|^2 + c_m\|u^+\|^2$ for all $u \in B_1E^m$. Therefore

$$\begin{aligned} I(u) &\ge \frac{1}{2}\|u^-\|^2 + c_3\|u^0\|^2 + c(m)\|u^+\|^\gamma - \frac{1}{2}\|u^+\|^2 \\ &= \frac{1}{2}\|u^-\|^2 + c_3\|u^0\|^2 + \left(c(m) - \frac{1}{2}\|u^+\|^{2-\gamma}\right)\|u^+\|^\gamma \\ &\ge \alpha_m\|u\|^2 \end{aligned}$$

if $u \in E^m$ with $\|u\| \le r_m := \min\{1, c(m)^{1/(2-\gamma)}\}$, where α_m is a positive constant depending on m. (i) is proved.

Let $u \in (E^{m-1})^{\perp}$. By Lemma 2.1, $|u|_\gamma \le c_\gamma m^{-1/2}\|u\|$ and so

$$I(u) \le c_1|u|_\gamma^\gamma - \frac{1}{2}\|u\|^2 \le c_2 m^{-\gamma/2}\|u\|^\gamma - \frac{1}{2}\|u\|^2$$
$$\le b_m := c_2\big(1-\frac{\gamma}{2}\big)\big(\frac{c_2\gamma}{m}\big)^{\gamma/(2-\gamma)}.$$

Clearly, $b_m \to 0$ as $m \to \infty$, and (ii) follows. □

In the following, fix arbitrarily $n \in \mathbb{N}$. Consider $I_l := I|_{E_{-ln}}$, the restriction of I onto E_{-ln} for all $l \in \mathbb{N}$. A standard verification shows that I and I_l satisfy the $(PS)_c$ condition, and I also satisfies the $(PS)_c^*$ condition for all $c \in \mathbb{R}$. Let H^l denote the set of all homeomorphisms $\lambda : E_{-ln} \to E_{-ln}$ satisfying

($\lambda 1$) λ is equivariant, i.e., $\lambda(T_\theta u) = T_\theta \lambda(u)$ for all $(\theta, u) \in [0, 2\pi) \times E_{-ln}$;

($\lambda 2$) $\lambda(0) = 0$;

($\lambda 3$) for any compact set K in a finite dimensional invariant space $Y \subset E_{-ln}$ and an $\varepsilon > 0$ there is a finite dimensional invariant space $Z \subset E_{-ln}$ with $Y \subset Z$ and an equivariant homeomorphism $\tilde\lambda : Z \to Z$ such that $\|\tilde\lambda(u) - \lambda(u)\| \le \varepsilon$ for all $u \in K$.

Let $(I_l)_a := \{u \in E_{-ln} :\ I(u) \ge a\}$ and $K_l^a := \{u \in E_{-ln} :\ I(u) = a \text{ and } I_l'(u) = 0\}$ for $a \in \mathbb{R}$. We recall the following standard result of deformation (cf. [8]).

LEMMA 5.2. *For any $0 \ne c \in \mathbb{R}$, if $\mathcal{N}$ is a neighborhood of K_l^c, then there exist $\bar\varepsilon > \varepsilon > 0$ and $\eta \in \mathcal{C}([0,1] \times E_{-ln}, E_{-ln})$ such that*

(a) $\eta_t \in H^l$ *for all* $t \in [0,1]$;

(b) $\eta_0(u) = u$ *for all* $u \in E_{-ln}$;

(c) $\eta_t(u) = u$ *if* $u \notin I^{-1}([c-\bar\varepsilon, c+\bar\varepsilon])$;

(d) $I(\eta_t(u)) \ge I(u)$ *for all* $(t,u) \in [0,1] \times E_{-ln}$;

(e) $\eta_1((I_l)_{c-\varepsilon} \backslash \mathcal{N}) \subset (I_l)_{c+\varepsilon}$;

(f) *if* $K_c^l = \emptyset, \eta_1((I_l)_{c-\varepsilon}) \subset (I_l)_{c+\varepsilon}$.

Set

$$\tilde E_{-ln} := \big(\oplus_{j=1}^{ln} E(-j)\big) \oplus E^+,$$
$$W_n^l := \big(\oplus_{j=1}^{l} E(-jn)\big) \oplus \overline{\oplus_{j=1}^{\infty} E(jn)},$$
$$\Sigma_n^l := \big\{A \subset E_{-ln} :\ A \text{ is closed, invariant and } A \subset W_n^l \oplus E(0)\big\}.$$

Let Σ be the set of all closed invariant subsets of E and let gen: $\Sigma \to \mathbb{N} \cup \{0, \infty\}$ be the S^1-index, that is, for $A \in \Sigma$

$$\operatorname{gen}(A) = \min\Big\{k \in \{0\} \cup \mathbb{N} :\ \begin{matrix}\text{there are } \phi \in \mathcal{C}(A, \mathbb{C}^k \backslash \{0\}) \text{ and } n \in \{0\} \cup \mathbb{N} \text{ such} \\ \text{that } \phi(T_\theta u) = e^{in\theta}\phi(u) \text{ for all } (u, e^{i\theta}) \in A \times S^1\end{matrix}\Big\}.$$

Define for $A \in \Sigma \cap E_{-ln}$

$$\operatorname{gen}_l(A) = \inf_{\lambda \in H^l} \operatorname{gen}\big(\lambda(A) \cap \tilde E_{-ln}\big)$$

and set

$$\Gamma_n^l := \{A \in \Sigma_n^l :\ \operatorname{gen}_l(A) \ge (l+1)N\}.$$

It is not difficult to check the following (cf. [8, Propositions 2.2(vii) and 2.11(i)]):

$$\eta(A) \in \Gamma_n^l \quad \text{for all } A \in \Gamma_n^l \text{ and } \eta \in H^l. \tag{5.6}$$

Since for $A \in \Gamma_n^l$, $(l+1)N \le \mathrm{gen}_l(A) \le \mathrm{gen}(A \cap \tilde{E}_{-ln})$ and

$$A \cap \tilde{E}_{-ln} \subset \big(\oplus_{j=1}^{l} E(-jn)\big) \oplus \big(\overline{\oplus_{j\in\mathbb{N}} E(jn)}\big),$$

one has $A \cap \big(\overline{\oplus_{j\in\mathbb{N}} E(jn)}\big) \neq \emptyset$. Thus

$$A \cap \big(E^{n-1}\big)^{\perp} \neq \emptyset \quad \text{for all } A \in \Gamma_n^l. \tag{5.7}$$

Moreover, by definition, the set

$$C_n^l := S_{r_n}\Big(\big(\oplus_{j=1}^{l} E(-jn)\big) \oplus E(n) \oplus E(0)\Big) \in \Gamma_n^l. \tag{5.8}$$

Now we are in a position to give the proof of Theorem 1.2(b).

Proof of Theorem 1.2(b). Fix $n \in \mathbb{N}$. For any $l \in \mathbb{N}$ we define the following minimax values for I_l:

$$b_n^l := \sup_{A\in\Gamma_n^l} \min_{u\in A} I(u).$$

Lemma 5.1(i) and (5.8) imply that $\alpha_n \le b_n^l$ for all $l \in \mathbb{N}$, and Lemma 5.1(ii) and (5.7) imply that $b_n^l \le \beta_n$ for all $l \in \mathbb{N}$, that is

$$\alpha_n \le b_n^l \le \beta_n \quad \text{for all } l \in \mathbb{N}. \tag{5.9}$$

By Lemma 5.2 and (5.6), (5.9), b_n^l is a critical value of I_l (cf. [9, 14, 17, 20]). Let $u_n^l \in E_{-ln}$ be such that

$$I(u_n^l) = b_n^l \quad \text{and} \quad I_l'(u_n^l) = 0.$$

Then by the $(PS)_c^*$ condition, along a subsequence as $l \to \infty$, $u_n^l \to u_n$ such that

$$\alpha_n \le I(u_n) \le \beta_n \quad \text{and} \quad I'(u_n) = 0. \tag{5.10}$$

Now by (5.5) and (5.10),

$$\begin{aligned}\beta_n \ge I(u_n) &= I(u_n) - \frac{1}{2} I'(u_n)u_n \\ &\ge \big(1 - \frac{\gamma}{2}\big) \int_0^{2\pi} \tilde{H}_{\xi\eta}(u_n),\end{aligned}$$

which, together with (5.4), implies

$$M|u_n|_\gamma^\gamma \le \gamma \int_0^{2\pi} \tilde{H}_{\xi\eta}(u_n) \le \frac{2\gamma\beta_n}{2-\gamma}. \tag{5.11}$$

By (G2) and the Hölder inequality $(1/\beta + 1/\beta' = 1, 1/\gamma + 1/\gamma' = 1)$

$$\begin{aligned}\|u_n^+\|^2 &= \int_0^{2\pi} \nabla \tilde{H}_{\xi\eta}(u_n)u_n^+ \\ &\le c_1|u_n^+|_{\beta'} \left(\int_{|u_n(t)|\le 1} |\nabla \tilde{H}_{\xi\eta}(u_n)|^\beta\right)^{1/\beta} + c_1|u_n^+|_\gamma \left(\int_{|u_n(t)|>1} |\nabla \tilde{H}_{\xi\eta}(u_n)|^{\gamma'}\right)^{1/\gamma'} \\ &\le c_2\|u_n^+\| \left[\left(\int_0^{2\pi} \tilde{H}_{\xi\eta}(u_n)\right)^{1/\beta} + \left(\int_0^{2\pi} \tilde{H}_{\xi\eta}(u_n)\right)^{1/\gamma'}\right]\end{aligned}$$

and so

$$\|u_n^+\| \le c_2\Big[\Big(\frac{2\beta_n}{2-\gamma}\Big)^{1/\beta} + \Big(\frac{2\beta_n}{2-\gamma}\Big)^{1/\gamma'}\Big]. \tag{5.12}$$

Similarly

$$\|u_n^-\| \le c_2\Big[\Big(\frac{2\beta_n}{2-\gamma}\Big)^{1/\beta} + \Big(\frac{2\beta_n}{2-\gamma}\Big)^{1/\gamma'}\Big]. \tag{5.13}$$

Equations (5.11)–(5.13) then yield

$$\|u_n\| \le c_3\Big[\Big(\frac{2\beta_n}{2-\gamma}\Big)^{1/\beta} + \Big(\frac{2\beta_n}{2-\gamma}\Big)^{1/\gamma'} + \Big(\frac{2\beta_n}{2-\gamma}\Big)^{1/\gamma}\Big]. \tag{5.14}$$

Since $\beta_n \to 0$ as $n \to \infty$, it follows from (5.14) that

$$\|u_n\| \to 0 \quad \text{as } n \to \infty. \tag{5.15}$$

Furthermore, since $I'(u_n) = 0, u_n$ solves

$$\dot z = \mathcal{J}\nabla \tilde H_{\xi\eta}(z),$$

and we obtain

$$\begin{aligned}\int_0^{2\pi} |\dot u_n|^2 &= \int_0^{2\pi} |\nabla \tilde H_{\xi\eta}(u_n)|^2 \\ &\le c_1 \int_0^{2\pi} \Big(\tilde H_{\xi\eta}(u_n)^{2/\gamma'} + \tilde H_{\xi\eta}(u_n)^{2/\beta}\Big) \\ &\le c_2 \left[\Big(\int_0^{2\pi} \tilde H_{\xi\eta}(u_n)\Big)^{2/\beta} + \Big(\int_0^{2\pi} \tilde H_{\xi\eta}(u_n)\Big)^{2/\gamma'}\right],\end{aligned}$$

which, jointly with (5.11), implies

$$|\dot u_n|_2 \to 0 \quad \text{as } n \to \infty. \tag{5.16}$$

Now (5.15) and (5.16) give the conclusion that

$$\|u_n\|_{W^{1,2}} \to 0 \quad \text{as } n \to \infty$$

and therefore, for n large, u_n solves ($\mathrm{HS}_{\xi\eta}$).

Finally we show that $u_n(t) \ne 0$ for all t. Indeed, if $u_n(t_0) = 0$ for some $t_0 \in \mathbb{R}$, then since $H_{\xi\eta}$ is independent t, one has

$$M|u_n(t)|^\gamma \le H_{\xi\eta}(u_n(t)) \equiv H_{\xi\eta}(u_n(t_0)) = 0,$$

i.e., $u_n(t) \equiv 0$, a contradiction.

The proof is complete. □

REFERENCES

[1] A. Ambrosetti, J. G. Azorero, and I. Peral, *Multiplicity results for some nonlinear elliptic equations*, J. Funct. Anal., 137 (1996), pp. 219–242.

[2] A. Ambrosetti, H. Brézis, and G. Cerami, *Combined effects of concave and convex nonlinearities in some elliptic problems*, J. Funct. Anal., 122 (1994), pp. 519–543.

[3] A. Bahri and H. Berestycki, *Forced vibrations of superquadratic Hamiltonian systems*, Acta Math., 152 (1984), pp. 143–197.

[4] T. Bartsch and M. Willem, *On an elliptic equation with concave and convex nonlinearities*, Proc. Amer. Math. Soc., 123 (1995), pp. 3555–3561.

[5] T. Bartsch and M. Willem, *Periodic solutions of nonautonomous Hamiltonian systems with symmetries*, J. Reine Angew. Math., 451 (1994), pp. 149–159.

[6] T. Bartsch and Y. H. Ding, *Periodic solutions of superlinear beam and membrane equations with perturbations from symmetry*, Nonlinear Anal., to appear.

[7] T. Bartsch, Y. H. Ding, and C. Lee, *Periodic solutions of a wave equation with concave and convex nonlinearities*, J. Differential Equations, 153 (1999), pp. 121–141.

[8] V. Benci, *A geometrical index for the group S^1 and some applications to the study of periodic solutions of ordinary differential equations*, Comm. Pure Appl. Math., 34 (1981), pp. 393–432.

[9] K. C. Chang, *Infinite Dimensional Morse Theory and Multiple Solution Problems*, Birkhäuser, Boston, Basel, Berlin, 1993.

[10] Y. H. Ding and M. Girardi, *Periodic solutions for a class of symmetric and subquadratic Hamiltonian systems*, Math. Comput. Modelling, 23 (1996), pp. 59–71.

[11] E. R. Fadell, S. Y. Husseini, and P. H. Rabinowitz, *Borsuk-Ulam theorems for arbitrary S^1 actions and applications*, Trans. Amer. Math. Soc., 274 (1982), pp. 345–360.

[12] Y. Long, *Periodic solutions of superquadratic Hamiltonian systems with bounded forcing terms*, Math. Z., 203 (1990), pp. 453–467.

[13] Y. Long, *Periodic solutions of perturbed superquadratic Hamiltonian systems*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 17 (1990), pp. 35–77.

[14] J. Mawhin and M. Willem, *Critical Point Theory and Hamiltonian Systems*, Appl. Math. Sci. 74, Springer-Verlag, New York, 1989.

[15] P. H. Rabinowitz, *Periodic solutions of Hamiltonian system*, Comm. Pure Appl. Math., 31 (1978), pp. 157–184.

[16] P. H. Rabinowitz, *Periodic solutions of large norm of Hamiltonian systems*, J. Differential Equations, 50 (1983), pp. 33–48.

[17] P. H. Rabinowitz, *Minimax Methods in Critical Point Theory with Applications to Differential Equations*, CBMS Reg. Conf. Ser. Math. 65, AMS, Providence, RI, 1986.

[18] K. Tanaka, *Infinitely many periodic solutions for the equation: $u_{tt} - u_{xx} \pm |u|^{p-1}u = f(x,t)$*, II, Trans. Amer. Math. Soc., 307 (1988), pp. 615–645.

[19] K. Tanaka, *Multiple periodic solutions of a superlinear forced wave equation*, Ann. Mat. Pura Appl. (4), 162 (1992), pp. 43–76.

[20] M. Willem, *Minimax Theorems*, Birkhäuser, Boston, Basel, Berlin, 1996.

SIAM J. MATH. ANAL.
Vol. 32, No. 3, pp. 572–587

GLOBAL EXISTENCE AND RELAXATION LIMIT FOR SMOOTH SOLUTIONS TO THE EULER–POISSON MODEL FOR SEMICONDUCTORS*

G. ALÌ†, D. BINI†, AND S. RIONERO‡

Abstract. We establish the global existence of smooth solutions of the Cauchy problem for the one-dimensional Euler–Poisson model for semiconductors, under the assumption that the initial data are perturbations of a stationary solution of the drift-diffusion equations. The resulting evolutionary solutions converge asymptotically in time to the unperturbed state.

Key words. Euler–Poisson, semiconductors, asymptotic behavior, smooth solutions

AMS subject classifications. 35L65, 76X05, 35M10

PII. S0036141099355174

1. Introduction. In this paper, we study the global existence and the asymptotic behavior of smooth solutions of the initial value problem for the Euler–Poisson (or hydrodynamic) model for semiconductors in one space dimension. In scaled variables, let n, u, T, and E denote the electron number density, the electron velocity, the electron temperature, and the electric field, respectively. The (nondimensional) Euler–Poisson model consists of a hydrodynamic part,

$$n_t + (nu)_x = 0, \tag{1.1}$$

$$(nu)_t + \left(nu^2 + nT\right)_x = nE - \frac{nu}{\tau}, \tag{1.2}$$

$$\left(\frac{nu^2}{2} + \frac{nT}{\gamma - 1}\right)_t + \left[\left(\frac{nu^2}{2} + \frac{\gamma nT}{\gamma - 1}\right) u - \kappa T_x\right]_x = nuE - \frac{1}{\sigma}\left(\frac{nu^2}{2} + \frac{n(T - T_L)}{\gamma - 1}\right), \tag{1.3}$$

supplemented by the Poisson equation

$$E_x = n - b. \tag{1.4}$$

Here, the positive constants γ ($\gamma > 1$), τ, and σ are the adiabatic exponent, the (scaled) momentum relaxation time, and the (scaled) energy relaxation time, respectively. The coefficient $\kappa = \tau\kappa'$ is the heat conductivity. In this paper, we will consider the case $\kappa = 0$, so that the hydrodynamic part of the model is hyperbolic. The functions $T_L(x)$ and $b(x)$ are the lattice temperature and the doping profile, respectively. For simplicity, we assume that the lattice temperature is constant. Then, we can set $T_L = 1$. We assume that the doping profile satisfies the conditions

$$b(x) \in C^2(\mathbb{R}), \tag{1.5}$$

*Received by the editors June 21, 1999; accepted for publication (in revised form) April 14, 2000; published electronically October 11, 2000.
http://www.siam.org/journals/sima/32-3/35517.html
†Institute for Applied Mathematics, CNR, via P. Castellino 111, I-80131 Napoli, Italy (ali@iamna.iam.na.cnr.it, binid@icra.it).
‡Dipartimento di Matematica, Università degli Studi Federico II, via Cinthia, I-80126 Napoli, Italy (rionero@matna2.dma.unina.it).

$$b(x) > 0, \quad x \in \mathbb{R}, \qquad \lim_{x \to \pm\infty} b(x) = b^{\pm} > 0. \tag{1.6}$$

The condition (1.5) could be weakened, for instance, assuming that there exists a function $B(x) \in C^2(\mathbb{R})$ such that

$$B(x) > 0, \quad B'(x) \in L^1(\mathbb{R}) \cap H^1(\mathbb{R}), \quad b(x) - B(x) \in H^1(\mathbb{R}).$$

The hydrodynamic model (1.1)–(1.4) has been introduced in [4] and [5], to describe electron flow when the transport of energy in the semiconductor plays a crucial role, as in submicron devices or in the occurrence of high field phenomena [16]. A careful discussion of the physical validity of this model can be found in [3].

Now we consider the transformation

$$t = \frac{t'}{\tau}, \qquad u = \tau u'. \tag{1.7}$$

Skipping the prime, the transformed variables satisfy

$$n_t + (nu)_x = 0, \tag{1.8}$$

$$(\alpha n u)_t + \left(\alpha n u^2 + nT\right)_x = nE - nu, \tag{1.9}$$

$$\left(\frac{\alpha}{2} n u^2 + \frac{nT}{\gamma - 1}\right)_t + \left(\frac{\alpha}{2} n u^3 + \frac{\gamma n T}{\gamma - 1} u - \kappa' T_x\right)_x = nuE - \frac{1}{\beta}\left(\frac{\alpha}{2} n u^2 + \frac{n(T-1)}{\gamma - 1}\right), \tag{1.10}$$

$$E_x = n - b, \tag{1.11}$$

where

$$\alpha = \tau^2, \quad \beta = \tau\sigma.$$

In terms of the original physical variables, we have

$$\alpha = \frac{k_B T_L}{m_e \bar{x}^2} \tau_p^2, \quad \beta = \frac{k_B T_L}{m_e \bar{x}^2} \tau_p \tau_w,$$

where k_B is the Boltzmann constant, m_e is the electron mass, τ_p is the momentum relaxation time, τ_w is the energy relaxation time, and $\bar{x}$ is a characteristic length scale. Typically, for an n^+-n-n^+ channel in a MOSFET [8], α is of order 10^{-2} and β of order 1.

We treat the positive numbers α and β as parameters. If α, $\beta \neq 0$, the system (1.8)–(1.11) is perfectly equivalent to the original system (1.1)–(1.4). A preliminary numerical and theoretical study of this model with nonzero heat conductivity $\kappa = \tau\kappa_0 nT$, where κ_0 is a positive constant, can be found in [8]. More recently, for the same model (with constant $\kappa > 0$), the global existence of smooth solutions, in a bounded domain and for small initial data, has been proved in [6] under the assumption that the doping profile is close enough to a constant function. If $\kappa' = 0$, (1.8)–(1.11) constitute a hyperbolic-elliptic system. Some analytic results are known for the related isentropic hydrodynamic model. In particular, the existence of weak solutions has been proved in [13] and [14], and the global existence of smooth solutions for small initial data has been proved in [11].

If $\alpha = 0$ and $\beta \neq 0$, we obtain a parabolic-elliptic system, known as the energy-transport model. The hydrodynamic part of this system is strongly parabolic if $\kappa' > 0$

and weakly parabolic if $\kappa' = 0$. For the time-dependent energy-transport model, no rigorous analytic results are known. The existence and uniqueness of stationary solutions for a general class of energy-transport models has been established in [7].

If $\alpha = \beta = 0$, we obtain the well-known time-dependent drift-diffusion model (see [16]).

The formal limits of (1.8)–(1.11) as α or β tend to zero have been obtained first in [2] and [1]. In some cases, this formal procedure can be rigorously justified. For $\kappa' = 0$, singular relaxation limit results have been obtained in [14] for the isentropic model and in [9] for the full system. In particular, in the last paper the authors postulate the existence of a time-dependent solution $(\bar{n}, \bar{u}, \bar{T}, \bar{E})(x,t)$ of the original system (1.1)–(1.4), satisfying the uniform bound

$$||(\bar{n}, \bar{u}, \bar{T})||_{L^\infty(\mathbb{R}\times\mathbb{R}_+)} \le \bar{C}, \tag{1.12}$$

where the positive constant $\bar{C}$ is independent on τ. Then, as τ tends to zero with β constant, the function $(n,u,T,E)(x,t) = (\bar{n}, \frac{1}{\tau}\bar{u}, \bar{T}, \bar{E})(x, \frac{1}{\tau}t)$ tends to a weak solution of (1.8)–(1.11) with $\alpha = 0$.

Also, we mention the singular relaxation limit result in [6], connecting the Euler–Poisson system (1.1)–(1.4), with $\kappa > 0$, to the drift-diffusion system as τ and σ tend to zero, with $0 \le 2\sigma - \tau \le M\sqrt{\tau\sigma}$.

In this paper, we are mainly interested in studying the global existence (in time) of solutions of the initial value problem for (1.8)–(1.11), with $\beta > \alpha \ge 0$ and $\kappa' = 0$.

Generally speaking, as time increases, we expect the solutions of (1.8)–(1.11) to approach the solutions of the corresponding stationary system. In particular, from (1.8), any stationary solution must satisfy

$$nu = j = \text{constant}. \tag{1.13}$$

If $j = 0$, the stationary solution reduces to $(n,u,T,E) = (\mathcal{N}, 0, 1, \mathcal{E})$, where $\mathcal{N}$ and $\mathcal{E}$ satisfy the stationary drift-diffusion equations

$$\begin{aligned} &\mathcal{N}_x = \mathcal{N}\mathcal{E}, \\ &\mathcal{E}_x = \mathcal{N} - b. \end{aligned} \tag{1.14}$$

In [11], the authors prove the existence and uniqueness of solutions to a slightly more general system than (1.14), satisfying the conditions

$$\mathcal{N}(x) - b(x) \in H^1(\mathbb{R}), \qquad \lim_{x\to-\infty} \mathcal{E}(x) = 0. \tag{1.15}$$

The function $\mathcal{N}(x)$ belongs to $C^2(\mathbb{R})$, and satisfies the estimate [11]

$$\inf_{x\in\mathbb{R}} b(x) \le \mathcal{N}(x) \le \sup_{x\in\mathbb{R}} b(x), \tag{1.16}$$

which ensures the strict positivity of $\mathcal{N}(x)$.

With these motivations, we assume that the initial data for (1.8)–(1.11),

$$\begin{aligned} &n(x,0) = n_0(x), \quad u(x,0) = u_0(x), \quad T(x,0) = T_0(x), \\ &E(x,0) = E_0(x) \equiv \int_{-\infty}^{x} (n_0(x') - b(x'))dx', \end{aligned} \tag{1.17}$$

are given as perturbations of the stationary solution $(\mathcal{N}(x), 0, 1, \mathcal{E}(x))$, satisfying (1.14). More precisely, we assume that the differences

$$n_0(x) - \mathcal{N}(x), \quad u_0(x) - 0, \quad T_0(x) - 1, \quad E_0(x) - \mathcal{E}(x)$$

belong to $H^2(\mathbb{R})$, and their H^2-norms are small enough. Also, we assume the conditions

$$\begin{aligned} &n(x,t) - b(x) \in H^1(\mathbb{R}), \\ &\lim_{x\to-\infty} E(x,t) = \lim_{x\to-\infty} u(x,t) = 0 \quad \forall t \in (0,+\infty). \end{aligned} \tag{1.18}$$

The only requirement for the unperturbed state is that

$$||\mathcal{E}||_{C^2} < +\infty. \tag{1.19}$$

Under these assumptions, we will show that the solution of the initial value problem (1.8)–(1.11) exists uniquely and globally in time and that it is a classical solution for $t > 0$. Moreover, it decays exponentially in the H^2-norm to the stationary solution, according to the estimate

$$\begin{aligned} &||(n - \mathcal{N}, \tau u, T - 1, E - \mathcal{E})(\cdot, t)||^2_{H^2} \\ &\leq K' e^{-Ct} ||(n - \mathcal{N}, \tau u, T - 1, E - \mathcal{E})(\cdot, 0)||^2_{H^2}, \end{aligned} \tag{1.20}$$

with K' and C positive constants independent on τ.

The a priori estimate (1.20) yields a remarkable consequence. Since (1.20) is valid also for $\alpha \equiv \tau^2 = 0$, it is possible to extend our global existence result to the energy-transport model. More precisely, let $(n^\alpha, u^\alpha, T^\alpha, E^\alpha)(x,t)$ be a solution of the system (1.8)–(1.11). Then, using (1.20) and the independence on α of the constants K' and C, there exist some limit functions $(\bar{n}, \bar{T}, \bar{E})(x,t)$ to which $(n^\alpha, T^\alpha, E^\alpha)(x,t)$ converge as α tends to zero. The limit solution satisfies the energy-transport model (1.8)–(1.11), with $\alpha = 0$. This result is closely related to the relaxation results proved in [9].

The proof of the estimate (1.20) is based on an energy method which is a modification of a method previously introduced in [15] for the compressible Navier–Stokes equations.

In [11], an a priori estimate of the kind (1.20) is derived for the isentropic model, under special assumptions on the doping profile. These assumptions amount to saying that the doping profile has to be close to a constant function, in a suitable norm. Using this condition it is possible to ensure the smallness of the C^2-norm of $\mathcal{E}$, which is required in the proof of the a priori estimate. The same kind of restriction on the doping profile is assumed in [6], with similar motivations. Unfortunately, this restriction is not compatible with the physical case of a doping profile with large gradient variations. This problem is not present in our approach, since the smallness of the C^2-norm of $\mathcal{E}$ is not required. The main idea is that, if (1.19) holds, we can control the effect of the first two derivatives of $\mathcal{E}$ on the solution of (1.8)–(1.11) by using a suitable weighted energy which combines the L^2-norm of the solution and the L^2-norm of its derivatives, as expounded in section 4.

The same method, with minor modifications, can be applied to the two- and three-dimensional version of (1.8)–(1.11). This extension will be the subject of a subsequent paper.

In section 2 we derive the perturbation equations around a given stationary solution, along with the appropriate initial data and boundary conditions. Then, in section 3, we discuss the positive definiteness of some functionals related to the definition of the energy. In the subsequent section, we apply the energy method in order to prove the a priori estimate (1.20). Finally, in section 5, we state and prove the consequent global existence result and asymptotic decay of the perturbation.

2. Perturbation equations. In this section, we derive the perturbation equations of system (1.8)–(1.11) around a steady solution $(n, u, T, E) = (\mathcal{N}, 0, 1, \mathcal{E})$, where $(\mathcal{N}, \mathcal{E})$ is the unique solution of (1.14), (1.15).

We consider the variable transformations

$$n = \mathcal{N} + v, \tag{2.1}$$

$$T = 1 + \theta, \tag{2.2}$$

$$E = \mathcal{E} + e, \tag{2.3}$$

with

$$|v| < \inf \mathcal{N}, \qquad |\theta| < 1. \tag{2.4}$$

The resulting equations for v, u, θ, and e are

$$v_t + [(\mathcal{N} + v)u]_x = 0, \tag{2.5}$$

$$\alpha u_t + \alpha u u_x + \frac{1}{\mathcal{N} + v}\left[(\mathcal{N} + v)(1 + \theta)\right]_x = e - u + \mathcal{E}, \tag{2.6}$$

$$\theta_t + u\theta_x + (\gamma - 1)(1 + \theta)u_x = au^2 - \frac{\theta}{\beta}, \tag{2.7}$$

$$e_x = v, \tag{2.8}$$

with

$$a = (\gamma - 1)\left(1 - \frac{\alpha}{2\beta}\right).$$

The initial data for v, u, θ, and e are given by

$$v(x, 0) = n_0(x) - \mathcal{N}(x), \qquad \theta(x, 0) = T_0(x) - 1, \tag{2.9}$$

$$u(x, 0) = u_0(x), \qquad e(x, 0) = \int_{-\infty}^{x} [n_0(y) - \mathcal{N}(y)]dy. \tag{2.10}$$

It is convenient to express the perturbations v and θ as

$$v = \mathcal{N}\nu,$$
$$\theta = (1 + \nu)^{\gamma - 1}(1 + s) - 1.$$

Inequalities (2.4) hold if

$$|\nu|, |s| < \frac{\log 2}{\gamma} \equiv \epsilon. \tag{2.11}$$

The functions ν, u, s, and e satisfy

$$\nu_t + [(1 + \nu)u]_x = -\mathcal{E}(1 + \nu)u, \tag{2.12}$$

$$\alpha u_t + \alpha u u_x + \frac{1}{1+\nu}\left[(1+\nu)^\gamma (1+s)\right]_x$$

$$= e - u - \mathcal{E}\left[(1+\nu)^{\gamma-1}(1+s) - 1\right], \tag{2.13}$$

$$s_t + u s_x = -\frac{(1+\nu)^{\gamma-1}(1+s) - 1}{\beta(1+\nu)^{\gamma-1}}$$

$$+\frac{a u^2}{(1+\nu)^{\gamma-1}} + (\gamma-1)\mathcal{E}(1+s)u, \tag{2.14}$$

$$e_x = \mathcal{N}\nu. \tag{2.15}$$

The initial data for ν and s are given by

$$\nu(x,0) = \frac{n_0(x)}{\mathcal{N}(x)} - 1, \quad s(x,0) = T_0(x)\left(\frac{\mathcal{N}(x)}{n_0(x)}\right)^{\gamma-1} - 1. \tag{2.16}$$

For regular solutions, the initial value problems (2.12)–(2.15), (2.16), (2.10), and (1.8)–(1.11), (1.17) are equivalent. Conditions (1.18) become

$$\begin{aligned} &\mathcal{N}(x)\nu(x,t) \in H^1(\mathbb{R}), \\ &\lim_{x\to-\infty} e(x,t) = \lim_{x\to-\infty} u(x,t) = 0 \quad \forall t \in (0,+\infty). \end{aligned} \tag{2.17}$$

Using (2.12), (2.17), and the choice of e_0 in (2.10), the constraint (2.15) can be replaced by the evolution equation

$$e_t + u e_x = -\mathcal{N}u. \tag{2.18}$$

Equations (2.12)–(2.14), (2.18) constitute a quasi-linear system of partial differential equations. It can be written in the form

$$I_\alpha \partial_t U + A_\alpha(U)\partial_x U = S_\alpha(U,x), \tag{2.19}$$

with $U = (\nu, u, s, e)$ and

$$I_\alpha = \operatorname{diag}(1, \alpha, 1, 1), \tag{2.20}$$

$$A_\alpha(U) = \begin{pmatrix} u & 1+\nu & 0 & 0 \\ \gamma(1+\nu)^{\gamma-2}(1+s) & \alpha u & (1+\nu)^{\gamma-1} & 0 \\ 0 & 0 & u & 0 \\ 0 & 0 & 0 & u \end{pmatrix}, \tag{2.21}$$

$$S_\alpha(U,x) = \begin{pmatrix} -\mathcal{E}(1+\nu)u \\ e - u - \mathcal{E}\left[(1+\nu)^{\gamma-1}(1+s) - 1\right] \\ -\frac{(1+\nu)^{\gamma-1}(1+s) - 1 - a\beta u^2}{\beta(1+\nu)^{\gamma-1}} + (\gamma-1)\mathcal{E}(1+s)u \\ -\mathcal{N}u \end{pmatrix}. \tag{2.22}$$

3. Positive definiteness of some functionals of Liapunov type. This section deals with the positive definiteness of some functionals of Liapunov type. The main results are summarized in the final lemma. They will be used in the subsequent section.

We introduce the energy densities

$$\begin{aligned}\mathcal{H}_0 &= \frac{\alpha}{2}\mathcal{N}(1+\nu)u^2 - \frac{\lambda_0}{\beta}\alpha(1+\nu)ue + \frac{1}{2}\left(1+\frac{\lambda_0}{\beta\mathcal{N}}\right)e^2 \\ &+\frac{\mathcal{N}}{\gamma-1}\left[(1+\nu)^\gamma - 1 - \gamma\nu\right](1+s) + \mathcal{N}\nu s + \frac{\mathcal{N}(1+\nu)s^2}{2(\gamma-1)}, \qquad (3.1)\\ \mathcal{H}_i &= \frac{\alpha}{2}\mathcal{N}(1+\nu)(\partial_x^i u)^2 - \frac{\lambda_i}{\beta}\alpha(1+\nu)\partial_x^i u\partial_x^i e \\ &+\frac{1}{2}\left(1+\frac{\lambda_i}{\beta\mathcal{N}}\right)(\partial_x^i e)^2 + \frac{\gamma}{2}\mathcal{N}(1+s)(1+\nu)^{\gamma-2}(\partial_x^i\nu)^2 \\ &+\mathcal{N}(1+\nu)^{\gamma-1}\partial_x^i\nu\partial_x^i s + \frac{\mathcal{N}(1+\nu)^\gamma(\partial_x^i s)^2}{2(\gamma-1)(1+s)}, \quad i=1,2. \qquad (3.2)\end{aligned}$$

Here, λ_0 and λ_i are positive constants. With an appropriate choice of λ_0 and λ_i, and if $|\nu|$ and $|s|$ are small enough, $\mathcal{H}_0$ and $\mathcal{H}_i$ are positive semidefinite, as asserted by the following lemma.

LEMMA 3.1. *If $0 \le \alpha < \beta$, $\lambda_i < 1/2$, and $|\nu|, |s| \le \epsilon'$ for some constant $\epsilon' < 1$, then there exist positive constants $k_{\mathcal{H}}$, $K_{\mathcal{H}}$ (depending on γ, $||\mathcal{N}||_{C^0}$, and ϵ', not depending on α) such that $\mathcal{H}_i$ ($i = 0, 1, 2$) satisfies*

$$k_{\mathcal{H}}|\partial_x^i(\nu, \tau u, s, e)| \le \mathcal{H}_i \le K_{\mathcal{H}}|\partial_x^i(\nu, \tau u, s, e)|,$$

where $\tau = \sqrt{\alpha}$.

Proof. We can decompose the energy densities as

$$\begin{aligned}\mathcal{H}_0 &= g_0(\nu, s) + h(\tau u, e), \\ \mathcal{H}_i &= g(\partial_x^i\nu, \partial_x^i s) + h(\tau\partial_x^i u, \partial_x^i e), \quad i=1,2,\end{aligned} \qquad (3.3)$$

where g_0, g, and h are the quadratic forms

$$\begin{aligned}g_0(X,Y) &= \mathcal{N}\left\{a(\nu)\frac{\gamma}{2}X^2 + b(\nu)XY + \frac{(1+\nu)}{2(\gamma-1)}Y^2\right\}, \\ g(X,Y) &= \mathcal{N}(1+\nu)^{\gamma-1}\left\{\left(\frac{1+s}{1+\nu}\right)\frac{\gamma}{2}X^2 + XY + \left(\frac{1+\nu}{1+s}\right)\frac{Y^2}{2(\gamma-1)}\right\}, \\ h(X,Y) &= \frac{\mathcal{N}}{2}(1+\nu)X^2 - \frac{\lambda_i}{\beta}\tau(1+\nu)XY + \frac{1}{2}\left(1+\frac{\lambda_i}{\beta\mathcal{N}}\right)Y^2.\end{aligned}$$

Here, the functions $a(\nu)$ and $b(\nu)$ are defined by

$$\begin{aligned}a(\nu) &= \frac{(1+\nu)^\gamma - 1 - \gamma\nu}{(\gamma/2)(\gamma-1)\nu^2} \quad \text{if } \nu \ne 0, \qquad a(0) = 1, \\ b(\nu) &= \frac{(1+\nu)^{\gamma-1} - 1}{(\gamma-1)\nu}(1+\nu) \quad \text{if } \nu \ne 0, \qquad b(0) = 1.\end{aligned}$$

First, we show that g_0 is positive definite for $|\nu| \le \epsilon' < 1$. It is simple to check that $a(\nu)$ and $b(\nu)$ are strictly positive and continuous functions for $|\nu| \le \epsilon'$. In particular, a is a decreasing function if $\gamma < 2$, a constant function if $\gamma = 2$, and an increasing function if $\gamma > 2$, and b is an increasing function. Then we have

$$\begin{aligned}&\frac{m_0}{2}\left(\gamma X^2 - 2|XY| + \frac{1}{\gamma-1}Y^2\right) \\ &\le g_0(X,Y) \le \frac{M_0}{2}\left(\gamma X^2 + 2|XY| + \frac{1}{\gamma-1}Y^2\right),\end{aligned}$$

where

$$m_0 = \inf \mathcal{N} \min\{a(-\epsilon'), a(\epsilon'), b(-\epsilon'), 1-\epsilon'\},$$
$$M_0 = \sup \mathcal{N} \max\{a(-\epsilon'), a(\epsilon'), b(\epsilon'), 1+\epsilon'\}.$$

The positive definiteness of g_0 follows immediately from the inequality

$$\frac{m_0}{2}\mu_0^-(X^2+Y^2) \le g_0(X,Y) \le \frac{M_0}{2}\mu_0^+(X^2+Y^2), \tag{3.4}$$

with

$$\mu_0^{\pm} = \frac{\gamma}{2} + \frac{1}{2(\gamma-1)} \pm \sqrt{\left(\frac{\gamma}{2} - \frac{1}{2(\gamma-1)}\right)^2 + 1}.$$

Next, we consider the quadratic form g. If $|\nu|, |s| \le \epsilon' < 1$, we have

$$\frac{m}{2}\left(\frac{\gamma}{r}X^2 + 2XY + \frac{r}{\gamma-1}Y^2\right)$$
$$\le g(X,Y) \le \frac{M}{2}\left(\frac{\gamma}{r}X^2 + 2XY + \frac{r}{\gamma-1}Y^2\right),$$

where

$$m = \inf \mathcal{N}(1-\epsilon')^{\gamma-1}, \quad M = \sup \mathcal{N}(1+\epsilon')^{\gamma-1}, \quad r = \frac{1+\nu}{1+s}.$$

The function r is bounded in the interval $(\frac{1-\epsilon'}{1+\epsilon'}, \frac{1+\epsilon'}{1-\epsilon'})$. The positive definiteness of g follows from the inequality

$$\frac{m}{2}\min \mu^-(r)(X^2+Y^2) \le g(X,Y) \le \frac{M}{2}\max \mu^+(r)(X^2+Y^2), \tag{3.5}$$

with

$$\mu^{\pm}(r) = \frac{\gamma}{2r} + \frac{r}{2(\gamma-1)} \pm \sqrt{\left(\frac{\gamma}{2r} - \frac{r}{2(\gamma-1)}\right)^2 + 1}.$$

Finally, the quadratic form h is positive definite if and only if

$$\alpha(1+\nu)\left(\frac{\lambda_i}{\beta}\right)^2 - \frac{\lambda_i}{\beta} - \mathcal{N} < 0,$$

which implies

$$\alpha\frac{\lambda_i}{\beta} < \frac{1+\sqrt{1+4\alpha\mathcal{N}(1+\nu)}}{2(1+\nu)}.$$

This inequality is always satisfied if $\lambda_i \le 1/2$ and $|\nu| < 1$. We have

$$\frac{1}{4}\inf \mu_h^-(X^2+Y^2) \le h(X,Y) \le \frac{1}{4}\sup \mu_h^+(X^2+Y^2), \tag{3.6}$$

where

$$\mu_h^{\pm} = 1 + \frac{\lambda_i}{\beta\mathcal{N}} + \mathcal{N}(1+\nu) \pm \sqrt{\left(1 + \frac{\lambda_i}{\beta\mathcal{N}} - \mathcal{N}(1+\nu)\right)^2 + \frac{\lambda_i^2}{\beta}(1+\nu)^2}.$$

The thesis of the lemma follows from (3.3), (3.4), (3.5), and (3.6). ∎

Now we define the functionals

$$\begin{aligned}\mathcal{D}_0 = \frac{\mathcal{N}}{\beta}\Big\{&(\beta - \alpha\lambda_0)u^2 + \frac{\lambda_0}{\mathcal{N}}e^2 - \frac{\lambda_0}{\mathcal{N}}\mathcal{E}(\gamma-1)e\nu - \frac{\lambda_0}{\mathcal{N}}\mathcal{E}es \\ &+(\gamma - 1 + \gamma\lambda_0)\nu^2 + (2+\lambda_0)\nu s + \frac{s^2}{\gamma-1}\Big\}, \end{aligned} \tag{3.7}$$

$$\begin{aligned}\mathcal{D}_i = \frac{\mathcal{N}}{\beta}\Big\{&(\beta - \alpha\lambda_i)(\partial_x^i u)^2 + \frac{\lambda_i}{\mathcal{N}}(\partial_x^i e)^2 \\ &+(\gamma - 1 + \gamma\lambda_i)(\partial_x^1\nu)^2 + (2+\lambda_i)\partial_x^1\nu\partial_x^1 s + \frac{(\partial_x^1 s)^2}{\gamma-1}\Big\}, \quad i = 1,2. \end{aligned} \tag{3.8}$$

We prove that the quadratic forms $\mathcal{D}_0$ and $\mathcal{D}_i$, are positive definite, with an appropriate choice of λ_0 and λ_i.

LEMMA 3.2. *If $0 \le \alpha < \beta$, $\lambda_i < 1$ $(i = 0,1,2)$, and*

$$\lambda_0 < \frac{4}{\gamma-1}\inf\left\{\frac{\mathcal{N}}{\mathcal{N}+\mathcal{E}^2}\right\}, \qquad \lambda_1, \lambda_2 < \frac{4}{\gamma-1},$$

then there exist positive constants $k_{\mathcal{D}}$, $K_{\mathcal{D}}$ (depending on γ and $||(\mathcal{N},\mathcal{E})||_{C^0}$, not depending on α) such that $\mathcal{D}_i$ $(i = 0,1,2)$ satisfy

$$k_{\mathcal{D}}|\partial_x^i U|^2 \le \mathcal{D}_i \le K_{\mathcal{D}}|\partial_x^i U|^2.$$

Proof. We can decompose the quadratic form $\beta\mathcal{D}_0/\mathcal{N}$ as

$$\beta\mathcal{D}_0/\mathcal{N}(u,e,\nu,s) = (\beta - \alpha\lambda_0)u^2 + \mathcal{D}'(e,\nu,s).$$

The coefficient of u^2 is positive, since $\lambda_0 < 1$ and

$$0 < \beta(1-\lambda_0) < \beta - \alpha\lambda_0 \le \beta. \tag{3.9}$$

The matrix associate to the quadratic form $\mathcal{D}'$ with respect to (e,ν,s) is

$$A = \begin{pmatrix} \lambda_0/\mathcal{N} & -\frac{1}{2}\lambda_0(\gamma-1)\mathcal{E}/\mathcal{N} & -\frac{1}{2}\lambda_0\mathcal{E}/\mathcal{N} \\ -\frac{1}{2}\lambda_0(\gamma-1)\mathcal{E}/\mathcal{N} & \gamma - 1 + \lambda_0\gamma & 1 + \lambda_0/2 \\ -\frac{1}{2}\lambda_0\mathcal{E}/\mathcal{N} & 1 + \lambda_0/2 & 1/(\gamma-1) \end{pmatrix}.$$

This matrix is definite positive if its determinant is positive together with the determinants of the following minors:

$$A_1 = \begin{pmatrix} \gamma - 1 + \lambda_0\gamma & 1 + \lambda_0/2 \\ 1 + \lambda_0/2 & 1/(\gamma-1) \end{pmatrix}, \qquad A_2 = \begin{pmatrix} 1/(\gamma-1) \end{pmatrix}.$$

Explicitly, these conditions amount to

$$\begin{aligned} &\frac{\lambda_0^2}{4\mathcal{N}}\left[\frac{4}{\gamma-1} - \lambda_0\left(\frac{\mathcal{N}+\mathcal{E}^2}{\mathcal{N}}\right)\right] > 0, \\ &\frac{\lambda_0}{4}\left(\frac{4}{\gamma-1} - \lambda_0\right) > 0, \\ &\frac{1}{\gamma-1} > 0. \end{aligned}$$

The previous inequalities are satisfied altogether if

$$0 < \lambda_0 < \frac{4}{\gamma - 1} \inf \left\{ \frac{\mathcal{N}}{\mathcal{N} + \mathcal{E}^2} \right\},$$

which holds by hypothesis. We can conclude that $\mathcal{D}_0$ is positive definite. Moreover, using (3.9), we can determine appropriate constants $k_{\mathcal{D}}$, $K_{\mathcal{D}}$, independent on α, such that the thesis holds. The positive definiteness of $\mathcal{D}_1$ and $\mathcal{D}_2$ follows immediately from the previous discussion, after setting $\mathcal{E} = 0$ and replacing λ_0 with λ_1 and λ_2, respectively. ∎

The following lemma follows immediately from Lemmas 3.1 and 3.2.

LEMMA 3.3. *If* $0 \le \alpha < \beta$, $|\nu|, |s| \le \epsilon' < 1$, *and*

$$\lambda_i < 1/2, \qquad \lambda_i < \frac{4}{\gamma - 1} \inf \left\{ \frac{\mathcal{N}}{\mathcal{N} + \mathcal{E}^2} \right\}, \qquad i = 0, 1, 2, \tag{3.10}$$

then there exist some positive constants $k_{\mathcal{H}}$, $K_{\mathcal{H}}$, $k_{\mathcal{D}}$, $K_{\mathcal{D}}$ *(dependent on* γ *and* ϵ'*) such that the functions* $\mathcal{H}_i$ *and* $\mathcal{D}_i$ *(*$i = 0, 1, 2$*) satisfy*

$$k_{\mathcal{H}} ||\partial_x^i(\nu, \tau u, s, e)||_{L^2}^2 \le \int \mathcal{H}_i dx \le K_{\mathcal{H}} ||\partial_x^i(\nu, \tau u, s, e)||_{L^2}^2, \tag{3.11}$$

$$k_{\mathcal{D}} ||\partial_x^i U||_{L^2}^2 \le \int \mathcal{D}_i dx \le K_{\mathcal{D}} ||\partial_x^i U||_{L^2}^2. \tag{3.12}$$

In particular, (3.11) *implies*

$$\int \mathcal{H}_i dx \le K'_{\mathcal{H}} ||\partial_x^i U||_{L^2}^2, \tag{3.13}$$

with

$$K'_{\mathcal{H}} = \max\{1, \beta\} K_{\mathcal{H}}.$$

4. A priori estimates. In this section, we establish a basic energy estimate for any given local solution of (2.12)–(2.15), with $|U| < \epsilon$. The constant ϵ is defined in (2.11).

For some fixed positive number T, we assume that a solution of (2.12)–(2.14), (2.18) exists for $t \in (0, T)$, and define

$$\mathcal{U}(T) = \sup_{0 \le t \le T} ||U(\cdot, t)||_{H^2}^2. \tag{4.1}$$

Using the standard Sobolev inequalities, there exists a positive constant $C_{\mathcal{U}}$ such that

$$\sup_{0 \le t \le T} ||U(\cdot, t)||_{C^1} \le C_{\mathcal{U}} \mathcal{U}(T). \tag{4.2}$$

We introduce the energy

$$\mathcal{W} = \int (\mathcal{H}_0 + \eta_1 \mathcal{H}_1 + \eta_2 \mathcal{H}_2)\, dx,$$

where the energy densities $\mathcal{H}_i$ are defined by (3.1), (3.2) and η_1, η_2 are positive constants to be determined.

LEMMA 4.1. *If $0 \le \alpha < \beta$, there exist positive constants ϵ', η_1, and η_2 such that if the solution is so small that $\mathcal{U}(T) \le \epsilon'$, the following a priori estimate holds for $t \in [0,T]$:*

$$\mathcal{W}(t) \le e^{-Ct}\mathcal{W}(0), \tag{4.3}$$

where the constant $C = k_{\mathcal{D}}/(4K'_{\mathcal{H}})$ is independent on α.

Proof. To begin with, we consider the function $\mathcal{H}_0$ defined by (3.1). A long but straightforward calculation shows that

$$\begin{aligned}
&\partial_t \mathcal{H}_0 = \{\cdots\}_x \\
&-\frac{\mathcal{N}}{\beta}(1+\nu)\left\{\left[\beta - \alpha\lambda_0 - (2\beta-\alpha)\frac{(1+\nu)^{\gamma-1}-1+s}{2(1+\nu)^{\gamma-1}}\right]u^2\right. \\
&+\frac{\lambda_0}{\mathcal{N}}e^2 + \left[\frac{((1+\nu)^{\gamma-1}-1)^2}{(\gamma-1)(1+\nu)^{\gamma-1}} + \frac{\lambda_0\nu}{1+\nu}((1+\nu)^{\gamma}-1)\right] \\
&\left.+\left[\frac{(1+\nu)^{2(\gamma-1)}-1}{(\gamma-1)(1+\nu)^{\gamma-1}} + \lambda_0(1+\nu)^{\gamma-1}\nu\right]s + \frac{s^2}{\gamma-1}\right\} \\
&+\mathcal{E}(1+\nu)\left[\frac{\lambda_0}{\beta}((1+\nu)^{\gamma-1}(1+s)-1)e + \alpha\frac{\lambda_0}{\beta}eu^2 + \mathcal{N}us^2\right] \\
&= \{\cdots\}_x - \mathcal{D}_0 + \mathcal{I}_0.
\end{aligned} \tag{4.4}$$

Here and in the following, the symbol $\{\cdots\}_x$ denotes the gradient of some unspecified function. This kind of term is not relevant, since it is going to vanish after integration with respect to x on the real line. In (4.4), $\mathcal{D}_0$ is the quadratic form defined by (3.7), and $\mathcal{I}_0$ collects all the remaining terms. For $|U| < \epsilon$, $\mathcal{I}_0$ is essentially cubic in the dependent variables, meaning that

$$|\mathcal{I}_0| \le c'_0|U|^3 \tag{4.5}$$

for some positive constant c'_0, independent on α. We note that c'_0 depends on the function $\mathcal{E}$, which appears in $\mathcal{I}_0$. Now, we integrate (4.4) with respect to x on the real line, and estimate the resulting right-hand side. From (4.5), we have

$$\int \mathcal{I}_0 dx \le c'_0||U||_{C^0}||U||^2_{L^2} \le c_0\mathcal{U}||U||^2_{L^2}, \tag{4.6}$$

with $c_0 \equiv c'_0 C_{\mathcal{U}}$. In conclusion, using (3.12), at zero order we find the estimate

$$\partial_t\left(\int \mathcal{H}_0 dx\right) = -\int \mathcal{D}_0 dx + \int \mathcal{I}_0 dx \le -(k_{\mathcal{D}} - c_0\mathcal{U})||U||^2_{L^2}. \tag{4.7}$$

Next, we consider the function $\mathcal{H}_1$, defined by (3.2). We find

$$\begin{aligned}
&\partial_t \mathcal{H}_1 = \{\cdots\}_x - \mathcal{D}_1 + \mathcal{I}_1 \\
&+\frac{\lambda_1}{\beta}\mathcal{E}\mathcal{N}\left[\alpha(1+\nu)^2 u u_x - (1+\nu)^{\gamma-1}(1+s)\nu\nu_x\right. \\
&\left.-\left(\frac{\beta}{\lambda_1} + \frac{1+\nu}{\mathcal{N}}\right)ue_x\right] + \frac{\lambda_1}{\beta}\mathcal{E}_x(1+\nu)\left[(1+\nu)^{\gamma-1}(1+s)-1\right]e_x \\
&-\mathcal{E}_x\mathcal{N}\left\{\left[(1+\nu)^{\gamma}(1+s)-1-\nu\right]u_x + (1+\nu)^{\gamma-1}(1+s)u\nu_x\right\}.
\end{aligned} \tag{4.8}$$

In (4.8), $\mathcal{D}_1$ is the quadratic form defined by (3.8) and $\mathcal{I}_1$ collects all the remaining terms not involving $\mathcal{E}$.

We integrate (4.8) with respect to x on the real line, and estimate the right-hand side for $|U| < \epsilon$. The function $\mathcal{I}_1$ is bounded by

$$|\mathcal{I}_1| \leq c_1'(|U| + |U_x|)|U_x|^2. \tag{4.9}$$

Then we have

$$\int \mathcal{I}_1 dx \leq c_1' ||U||_{C^1} ||U_x||_{L^2}^2 \leq c_1 \mathcal{U} ||U_x||_{L^2}^2, \tag{4.10}$$

with $c_1 \equiv c_1' C_{\mathcal{U}}$. For the remaining terms in (4.8) involving $\mathcal{E}$, we find

$$\begin{aligned} &\int \left\{ \frac{\lambda_1}{\beta} \mathcal{E}\mathcal{N} \left[\alpha(1+\nu)^2 u u_x - (1+\nu)^{\gamma-1}(1+s)\nu\nu_x \right.\right. \\ &\left. - \left(\frac{\beta}{\lambda_1} + \frac{1+\nu}{\mathcal{N}} \right) u e_x \right] + \frac{\lambda_1}{\beta} \mathcal{E}_x (1+\nu) \left[(1+\nu)^{\gamma-1}(1+s) - 1 \right] e_x \\ &\left. - \mathcal{E}_x \mathcal{N} \left[\left((1+\nu)^{\gamma}(1+s) - 1 - \nu \right) u_x + (1+\nu)^{\gamma-1}(1+s) u \nu_x \right] \right\} dx \\ &\leq 2 d_1 ||\mathcal{E}||_{C^1} ||U||_{L^2} ||U_x||_{L^2} \leq d_1 ||\mathcal{E}||_{C^1} \left(\frac{||U||_{L^2}^2}{\alpha_1} + \alpha_1 ||U_x||_{L^2}^2 \right). \end{aligned} \tag{4.11}$$

Here, d_1 is a positive constant, α_1 is a positive parameter that will be appropriately chosen later. In conclusion, using (3.12), at first order we find the estimate

$$\begin{aligned} \partial_t \left(\int \mathcal{H}_1 dx \right) &\leq \frac{d_1}{\alpha_1} ||\mathcal{E}||_{C^1} ||U||_{L^2}^2 \\ &- \left(k_{\mathcal{D}} - c_1 \mathcal{U} - \alpha_1 d_1 ||\mathcal{E}||_{C^1} \right) ||U_x||_{L^2}^2. \end{aligned} \tag{4.12}$$

Next, we consider the function $\mathcal{H}_2$, defined by (3.2). Derivating with respect to time, and using the identity

$$e_{xx} = \mathcal{N}\nu_x + \mathcal{E} e_x,$$

we can write

$$\partial_t \mathcal{H}_2 = \{\cdots\}_x - \mathcal{D}_2 + \mathcal{I}_2 + \text{terms containing } \mathcal{E},\ \mathcal{E}_x,\ \mathcal{E}_{xx}. \tag{4.13}$$

In (4.13), $\mathcal{D}_2$ is the quadratic form defined by (3.8), and $\mathcal{I}_2$ collects all the remaining terms not involving the function $\mathcal{E}$.

We integrate (4.13) with respect to x on the real line, and estimate the right-hand side for $|U| < \epsilon$ and $|\nu_x| < \epsilon$. We find that $\mathcal{I}_2$ is bounded by

$$|\mathcal{I}_2| \leq c_2' \left\{ (|U| + |U_x|)|U_{xx}|^2 + |U_x|^3 \right\}. \tag{4.14}$$

Then we have

$$\begin{aligned} \int \mathcal{I}_2 dx &\leq c_2' \left\{ ||U||_{C^1} ||U_{xx}||_{L^2}^2 + ||U||_{C^0} ||U_x||_{L^2}^2 \right\} \\ &\leq c_2 \mathcal{U} \left\{ ||U_{xx}||_{L^2}^2 + ||U_x||_{L^2}^2 \right\}, \end{aligned} \tag{4.15}$$

with $c_2 \equiv c_2' C_{\mathcal{U}}$.

The remaining terms involving $\mathcal{E}$ and its first two derivatives can be estimated as

$$\begin{aligned} &\int \{\text{terms containing } \mathcal{E},\, \mathcal{E}_x,\, \mathcal{E}_{xx}\}\, dx \\ &\leq d_2 ||\mathcal{E}||_{C^2} \left(\frac{||U||_{L^2}^2 + ||U_x||_{L^2}^2}{2\alpha_2} + \alpha_2 ||U_{xx}||_{L^2}^2 \right). \end{aligned} \tag{4.16}$$

Here, d_2 is a positive constant, α_2 is a positive parameter that will be appropriately chosen later. In conclusion, using (3.12), at second order we find the estimate

$$\begin{aligned} &\partial_t \left(\int \mathcal{H}_2 dx \right) \leq \frac{d_2}{2\alpha_2} ||\mathcal{E}||_{C^2} (||U||_{L^2}^2 + ||U_x||_{L^2}^2) \\ &+c_2 \mathcal{U} ||U_x||_{L^2}^2 - (k_{\mathcal{D}} - c_2 \mathcal{U} - \alpha_2 d_2 ||\mathcal{E}||_{C^2})\, ||U_{xx}||_{L^2}^2. \end{aligned} \tag{4.17}$$

Now we can estimate the energy $\mathcal{W}$ by using (4.7), (4.12), and (4.17). We find

$$\begin{aligned} &\partial_t \mathcal{W} \equiv \partial_t \left(\int \mathcal{H}_0 dx + \eta_1 \int \mathcal{H}_1 dx + \eta_2 \int \mathcal{H}_2 dx \right) \\ &\leq - \left(k_{\mathcal{D}} - c_0 \mathcal{U} - \frac{\eta_1 d_1}{\alpha_1} ||\mathcal{E}||_{C^1} - \frac{\eta_2 d_2}{2\alpha_2} ||\mathcal{E}||_{C^2} \right) ||U||_{L^2}^2 \\ &-\eta_1 \left(k_{\mathcal{D}} - \bar{c}_1 \mathcal{U} - \alpha_1 d_1 ||\mathcal{E}||_{C^1} - \frac{\eta_2 d_2}{2\eta_1 \alpha_2} ||\mathcal{E}||_{C^2} \right) ||U_x||_{L^2}^2 \\ &-\eta_2 \left(k_{\mathcal{D}} - c_2 \mathcal{U} - \alpha_2 d_2 ||\mathcal{E}||_{C^2} \right) ||U_{xx}||_{L^2}^2, \end{aligned} \tag{4.18}$$

with $\bar{c_1} = c_1 + (\eta_2/\eta_1) c_2$. We recall that $k_{\mathcal{D}}$, c_0, c_1, c_2, d_1, and d_2 are positive constants (independent on α), α_1, α_2, η_1, and η_2 are positive parameters to be chosen. We choose

$$\alpha_1 = \frac{k_{\mathcal{D}}}{4 d_1 ||\mathcal{E}||_{C^1}}, \qquad \alpha_2 = \frac{k_{\mathcal{D}}}{2 d_2 ||\mathcal{E}||_{C^2}}, \tag{4.19}$$

$$\eta_1 = \frac{2 k_{\mathcal{D}}^2}{k_{\mathcal{D}}^2 + 16 d_1^2 ||\mathcal{E}||_{C^1}^2}, \qquad \eta_2 = \frac{\eta_1 k_{\mathcal{D}}^2}{4 d_2^2 ||\mathcal{E}||_{C^2}^2}. \tag{4.20}$$

Substituting in (4.18), we obtain

$$\begin{aligned} &\partial_t \mathcal{W} \leq - (k_{\mathcal{D}}/2 - c_0 \mathcal{U})\, ||U||_{L^2}^2 \\ &-\eta_1 (k_{\mathcal{D}}/2 - \bar{c}_1 \mathcal{U})\, ||U_x||_{L^2}^2 - \eta_2 (k_{\mathcal{D}}/2 - c_2 \mathcal{U})\, ||U_{xx}||_{L^2}^2. \end{aligned} \tag{4.21}$$

If the solution satisfies

$$\mathcal{U}(T) \leq \epsilon' \equiv \min \left\{ \epsilon, \frac{k_{\mathcal{D}}}{4 \max\{c_0, \bar{c}_1, c_2\}} \right\}, \tag{4.22}$$

we have

$$\partial_t \mathcal{W} \leq -(k_{\mathcal{D}}/4) \left(||U||_{L^2}^2 + \eta_1 ||U_x||_{L^2}^2 + \eta_2 ||U_{xx}||_{L^2}^2 \right). \tag{4.23}$$

Then, using (3.13), we get the estimate

$$\partial_t \mathcal{W} \leq - \frac{k_{\mathcal{D}}}{4 K_{\mathcal{H}}'} \mathcal{W}. \tag{4.24}$$

The a priori estimate sought follows immediately from (4.24). □

LEMMA 4.2. *With the same assumptions of Lemma* 4.1, *the following a priori estimate holds for* $t \in [0, T]$*:*

$$||(\nu, \tau u, s, e)(\cdot, t)||^2_{H^2} \leq Ke^{-Ct}||(\nu, \tau u, s, e)(\cdot, 0)||^2_{H^2}, \tag{4.25}$$

where $\tau = \sqrt{\alpha}$*, and the positive constant* C *and* K *are given by*

$$C = \frac{k_{\mathcal{D}}}{4K_{\mathcal{H}} \max\{1, \beta\}}, \qquad K = \frac{K_{\mathcal{H}} \max\{1, \eta_1, \eta_2\}}{k_{\mathcal{H}} \min\{1, \eta_1, \eta_2\}}.$$

5. Global existence in time and asymptotic decay of the perturbation. In this section we state and prove the main theorem of the paper, Theorem 5.1, asserting the global existence in time of classical solutions of the perturbation equations (2.12)–(2.15) and their decay to the unperturbed state. We assume that the reader is familiar with the classical results on the local existence and continuation of regular solutions of quasi-linear hyperbolic systems (see [12]).

Let us consider the Euler–Poisson system (2.12)–(2.15). Using (2.12), (2.17), and the choice of e_0 in (2.10), we have shown in section 2 that the constraint (2.15) can be replaced by the evolution equation (2.18), obtaining the quasi-linear system (2.19). In particular, if $\alpha \neq 0$, (2.19) can be written in the form

$$\partial_t U + A(U)\partial_x U = S(U, x, t), \tag{5.1}$$

with $U = (\nu, u, s, e)$ in the state space $G = \{U : |\nu| < 1, |s| < 1\} \subseteq \mathbb{R}^4$, and

$$A = I_\alpha^{-1} A_\alpha, \qquad S = I_\alpha^{-1} S_\alpha.$$

The system (5.1) is hyperbolic and the matrix $A(U)$ satisfies the following property: for any U there is a positive definite symmetric matrix $\tilde{A}(U)$ smoothly varying with U, and a positive constant c, so that, $\forall U \in G_1$, $\bar{G}_1 \subset G$,

1. $cV \cdot V \leq (\tilde{A}(U)V) \cdot V \leq c^{-1} V \cdot V \ \forall V \in G$;
2. $\tilde{A}(U)A(U)$ is symmetric.

Specifically, the symmetry condition is satisfied with $G_1 = \{U : |\nu| < \epsilon, |s| < \epsilon\}$ and

$$\tilde{A}(U) = \begin{pmatrix} \gamma(1+\nu)^{\gamma-2}(1+s) & 0 & (1+\nu)^{\gamma-1} & 0 \\ 0 & \alpha(1+\nu) & 0 & 0 \\ (1+\nu)^{\gamma-1} & 0 & (1+\nu)^{\gamma}/(1+s) & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

We denote $U_0 = (\nu_0, u_0, s_0, e_0)$. Now, we are ready to state the global existence theorem announced at the beginning of this section.

THEOREM 5.1 (global existence and asymptotic decay). *If* $0 < \alpha < \beta$ *and* $U_0 \in H^2$*, then there is a positive constant* ϵ' *such that, if*

$$||U_0||_{H^2} < \epsilon',$$

the equations (2.12)–(2.15)*, with condition* (2.17)*, have a unique classical solution* $U(x, t) \in C^1(\mathbb{R} \times [0, \infty))$*. Furthermore,*

$$U \in C^0([0, \infty), H^2) \cap C^1([0, \infty), H^1)$$

and

$$||(\nu, \tau u, s, e)(\cdot, t)||^2_{H^2} \leq Ke^{-Ct}||(\nu, \tau u, s, e)(\cdot, 0)||^2_{H^2},$$

where K and C are positive constants, given by Lemma 4.2*, and $\tau = \sqrt{\alpha}$.*

Proof. Standard methods yield immediately the local H^2 existence of a unique classical solution to the initial value problem for (2.12)–(2.15), written in the form (5.1). Then, using Lemma 4.2, the local solution can be prolonged for all times and the estimate (4.25) holds globally. □

We remark that the constants in the estimate (4.25) are independent on α. Then, the estimate is also valid as α tends to zero, and the limit function is a solution of (2.12)–(2.15) with $\alpha = 0$. This result is better expressed in terms of the original variables $(n, u, T, E)(x, t)$ in (1.1)–(1.4). Using (1.7), we define the rescaled variables

$$(n', u', T', E')(x, t') = \left(n, \frac{1}{\tau}u, T, E\right)\left(x, \frac{1}{\tau}t'\right)$$

and, recalling (1.17), we consider the initial data

$$(n', u', T', E')(x, 0) = (n_0, u_0, T_0, E_0)(x).$$

Using Theorem 5.1, if the initial data for the primed variables are close enough to the steady solution $(\mathcal{N}, 0, 1, \mathcal{E})(x)$ in H^2, then

$$\begin{aligned} &||(n' - \mathcal{N}, \tau u', T' - 1, E' - \mathcal{E})(\cdot, t')||^2_{H^2} \\ &\leq K'e^{-Ct'}||(n_0 - \mathcal{N}, \tau u_0, T_0 - 1, E_0 - \mathcal{E})||^2_{H^2}, \end{aligned} \tag{5.2}$$

with K' and C positive constants independent on τ. This implies

$$\begin{aligned} &||(n - \mathcal{N}, u, T - 1, E - \mathcal{E})(\cdot, t)||^2_{H^2} \\ &\leq K'e^{-C\tau t}||(n_0 - \mathcal{N}, \tau u_0, T_0 - 1, E_0 - \mathcal{E})||^2_{H^2}. \end{aligned} \tag{5.3}$$

THEOREM 5.2 (relaxation). *Let $\beta > 0$. For any fixed $\alpha \equiv \tau^2 > 0$, let $(n, u, T, E) = (n^\alpha, u^\alpha, T^\alpha, E^\alpha)(x, t)$ be a global solution of* (1.1)–(1.4)*, satisfying* (5.3)*. Then, there exist some functions $(\hat{n}, \hat{T}, \hat{E})$ which are a smooth solution of* (1.8)–(1.11)*, with $\alpha = 0$, and such that, as α tends to zero,*

$$(n^\alpha, T^\alpha, E^\alpha)\left(x, \frac{1}{\tau}t\right) \to (\hat{n}, \hat{T}, \hat{E})(x, t) \quad \text{in } C([0, \infty); H^2).$$

Furthermore,

$$\begin{aligned} &||(\hat{n} - \mathcal{N}, \hat{T} - 1, \hat{E} - \mathcal{E})(\cdot, t)||^2_{H^2} \\ &\leq K'e^{-Ct}||(n - \mathcal{N}, T - 1, E - \mathcal{E})(\cdot, 0)||^2_{H^2}, \end{aligned} \tag{5.4}$$

where K' and C are positive constants.

Acknowledgments. Ministero dell'Università e della Ricerca Scientifica e Tecnocologica (M.U.R.S.T.) and the Gruppo Nazionale per la Fisica Matematica (G.N.F.M.) of the Italian Istituto Nazionale di Alta Matematica (I.N.D.A.M.) are acknowledged by the authors. G. Alì and D. Bini are indebted to Dott. R. Natalini for many stimulating discussions and to Prof. M. R. Occorsio for his continuous encouragement.

REFERENCES

[1] G. Alì, *Asymptotic Fluid Dynamic Models for Semiconductors*, Quaderno IAC 25/1995, Roma, 1995.

[2] G. Alì, P. Marcati, and R. Natalini, *Hydrodynamic models for semiconductors*, Z. Angew. Math. Mech., 76 supp. 2 (1996), pp. 301–304.

[3] A. M. Anile, *An extended thermodynamic framework for the hydrodynamical modeling of semiconductors*, in Mathematical Problems in Semiconductor Physics (Rome, 1993), Pitman Res. Notes Math. Ser. 340, P. A. Marcati, P. A. Markowich, and R. Natalini, eds., Longman Harlow, UK, 1995, pp. 3–41.

[4] K. Bløtekjær, *Transport equations for electrons in two-valley semiconductors*, IEEE Trans. Electron. Devices, 17 (1970), pp. 38–47.

[5] G. Baccarani and M. R. Wordeman, *An investigation of steady–state velocity overshoot effects in Si and GaAs devices*, Solid State Electr., 28 (1985), pp. 407–416.

[6] G. Q. Chen, J. W. Jerome, and B. Zhang, *Existence and the singular relaxation limit for the inviscid hydrodynamic energy model*, in Modelling and Computation for Applications in Mathematics, Science, and Engineering (Evanston, IL, 1996), Numer. Math. Sci. Comput., Oxford University Press, New York, 1998, pp. 189–215.

[7] P. Degond, S. Genieys, and A. Jüngel, *A steady-state system in nonequilibrium thermodynamics including thermal and electrical effects*, C. R. Acad. Sci. Paris Ser. I Math., 324 (1997), pp. 867–872.

[8] C. L. Gardner, J. W. Jerome, and D. J. Rose, *Numerical methods for the hydrodynamic device model: Subsonic flow*, IEEE Trans. Electron. Devices, 8 (1989), pp. 501–507.

[9] I. Gasser and R. Natalini, *The energy transport and the drift diffusion equations as relaxation limits of the hydrodynamic model for semiconductors*, Quart. Appl. Math., 57 (1999), pp. 269–282.

[10] H. Hattori and D. Li, *Global solutions of a high-dimensional system for Korteweg materials*, J. Math. Anal. Appl., 198 (1996), pp. 84–97.

[11] T. Luo, R. Natalini, and Z. Xin, *Large time behavior of the solutions to a hydrodynamic model for semiconductors*, SIAM J. Appl. Math., 59 (1998), pp. 810–830.

[12] A. Majda, *Compressible Fluid Flow and Systems of Conservation Laws in Several Space Variables*, Appl. Math. Sci. 53, Springer-Verlag, New York, Berlin, Heidelberg, Tokyo, 1984.

[13] P. Marcati and R. Natalini, *Weak solutions to a hydrodynamic model for semiconductors: The Cauchy problem*, Proc. Roy. Soc. Edinburgh Sect. A, 28 (1995), pp. 115–131.

[14] P. Marcati and R. Natalini, *Weak solutions to a hydrodynamic model for semiconductors and relaxation to the drift–diffusion equation*, Arch. Rational Mech. Anal., 129 (1995), pp. 129–145.

[15] A. Matsumura and T. Nishida, *The initial value problem for the equations of motion of viscous and heat-conductive gases*, J. Math. Kyoto Univ., 20 (1980), pp. 67–104.

[16] P. A. Markowich, C. A. Ringhofer, and C. Schmeiser, *Semiconductor Equations*, Springer-Verlag, Wien, New York, 1990.

SIAM J. MATH. ANAL.
Vol. 32, No. 3, pp. 588–615

OPTIMAL SCHAUDER ESTIMATES FOR PARABOLIC PROBLEMS WITH DATA MEASURABLE WITH RESPECT TO TIME*

LUCA LORENZI†

Abstract. We prove some optimal Schauder estimates for the solution to second-order parabolic equations with coefficients which are measurable with respect to time and Hölder continuous with respect to space variables in the strip $[0,T]\times\mathbb{R}^n$. We allow also polynomially or exponentially weighted Hölder norms.

Key words. linear parabolic equations in $\mathbb{R}^n$ with measurable coefficients, Cauchy problems, weighted spaces, Schauder estimates

AMS subject classifications. 35K15, 35K22, 35B45, 35B65, 35R05

PII. S0036141098342842

1. Introduction.

In this paper, we prove some sharp Schauder estimates and regularity properties for the parabolic second-order Cauchy problem

$$\begin{cases} D_t u(t,x) = \displaystyle\sum_{i,j=1}^{n} q_{i,j}(t,x) D_{i,j} u(t,x) + \sum_{i=1}^{n} b_i(t,x) D_i u(t,x) \\ \qquad + c(t,x)u(t,x) + f(t,x), \qquad (t,x)\in[0,T]\times\mathbb{R}^n, \\ u(0,x) = u_0(x), \qquad x\in\mathbb{R}^n. \end{cases} \tag{1.1}$$

Problem (1.1) has been studied in [7], [8], [9], where the coefficients $Q=[q_{i,j}]$, $B=[b_j]$, and c are bounded and continuous functions in $[0,T]\times\mathbb{R}^n$, Hölder continuous with respect to the space variables, with Hölder norms not depending on t. Moreover, $Q(t,x)$ is assumed to be a definite positive matrix, uniformly with respect to (t,x), and problem (1.1) is studied in the space of bounded and continuous functions in $[0,T]\times\mathbb{R}^n$.

In this paper, we extend the results of [7], [8], [9] to the case of less regular coefficients and of weighted Hölder norms. To be more precise, we analyze the case where the coefficients are bounded and measurable in $[0,T]\times\mathbb{R}^n$ and Hölder continuous in x, uniformly with respect to the variable t. Moreover, we study problem (1.1) in the context of the weighted space $UC_p(\mathbb{R}^n)$ choosing as p either the polynomial function, defined by $p(x)=1+|x|^{2m}$ $\forall x\in\mathbb{R}^n$ ($m\in\mathbb{N}\cup\{0\}$), or the exponential function, defined by $p(x)=\exp((1+|S^{1/2}x|^2)^{1/2})$, $\forall x\in\mathbb{R}^n$, S being any strictly positive and symmetric matrix. We assume also that $Q(t,x)$ is a strictly definite positive and symmetric matrix in $\mathcal{P}\times\mathbb{R}^n$, where $\mathcal{P}$ is a measurable set in $[0,T]$ such that $\mathcal{P}^c\cap[0,T]$ is negligible.

The problem of determining Schauder estimates for parabolic problems, when the coefficients are not continuous in time, has been treated by several authors. Brandt [2] deals with interior Hölder regularity with respect to the space variables, for a wide class of parabolic second-order operators with discontinuous coefficients, by means of a maximum principle and a perturbation argument. Then in [6], the results of [2] are

*Received by the editors June 30, 1998; accepted for publication (in revised form) May 9, 2000; published electronically October 20, 2000.
http://www.siam.org/journals/sima/32-3/34284.html
†Department of Mathematics, University of Parma, via M. D'Azeglio 85/A, I-43100 Parma, Italy (lorenzi@prmat.math.unipr.it).

extended by showing Hölder regularity results, also with respect to the time variable, for the solution to the second-order parabolic problem considered in [2]. Even if in [2] and [6] the authors consider also differential operators with coefficients and datum f that may be discontinuous in time, they are concerned with classical solutions since the proofs of their results are based, essentially, on the maximum principle in [5].

Other related papers concerning Schauder estimates for second-order initial-boundary problems are [10], [14], [3], [1], and [4]. Lieberman [10] is concerned with parabolic problems in bounded domains where the coefficients and the inhomogeneous term are Hölder continuous with respect to the space variables, measurable in time, and, possibly, unbounded near the boundary.

In [3], the authors are concerned with interior $W^{2,p}_{\mathrm{loc}}(\mathbb{R}^n)$ estimates for a solution to a second-order elliptic equation in nondivergence form assuming that the coefficients B and c vanish in $\mathbb{R}^n$ while Q does not depend on t and belongs to $VMO(\mathbb{R}^n)$. In [1], the authors provide interior and boundary estimates in $W^{1,2}_p$ for parabolic problems in the cylinder $\Omega\times(0,T)$, with a smooth enough lateral surface, assuming the second-order operator L to coincide with its principal part and having coefficients in VMO.

Finally, the paper [14] deals with an existence and uniqueness result for the classical solution to (1.1) in weighted spaces, where the weight p is a smooth function depending also on the time variable and increasing no more than polynomially and the coefficients Q, B, and $c<0$ are continuous $\mathbb{R}\times\mathbb{R}^n$. Before stating our main result, we give the following definition of solution to problem (1.1).

DEFINITION 1.1. *A function $u:[0,T]\times\mathbb{R}^n\to\mathbb{R}$ is called a solution to* (1.1) *if the following conditions are fulfilled:*

(i) *$u/p\in \mathrm{Lip}\,([0,T]\times\mathbb{R}^n)$, its first- and second-order space derivatives are continuous and bounded functions in $[0,T]\times\mathbb{R}^n$;*

(ii) *$u(0,x)=u_0(x)$ for any $x\in\mathbb{R}^n$;*

(iii) *there exists a negligible set $F\subset[0,T]\times\mathbb{R}^n$ such that $D_tu(t,x)=\mathcal{A}u(t,x)+f(t,x)$ for any $(t,x)\in([0,T]\times\mathbb{R}^n)\setminus F$. Moreover, for any $x\in\mathbb{R}^n$, the set $F(x)=\{t\in[0,T]:(t,x)\in F\}$ is measurable with measure T.*

THEOREM 1.2. *For any $u_0\in C^{2+\theta}_p(\mathbb{R}^n)$ and any measurable function $f:[0,T]\times\mathbb{R}^n\to\mathbb{R}$ such that $f(t,\cdot)\in C^\theta_p(\mathbb{R}^n)$, for any $t\in[0,T]$ and $\sup_{t\in[0,T]}\|f(t,\cdot)\|_{C^\theta_p(\mathbb{R}^n)}<+\infty$, there exists a unique function $u:[0,T]\times\mathbb{R}^n\to\mathbb{R}$ solution to problem* (1.1) *in the sense of Definition* 1.1 *belonging to $B([0,T];C^{2+\theta}_p(\mathbb{R}^n))$. Moreover, there exists a positive constant C, independent of (u,u_0,f), such that*

$$\sup_{t\in[0,T]}\|u(t,\cdot)\|_{C^{2+\theta}_p(\mathbb{R}^n)}\le C\left(\|u_0\|_{C^{2+\theta}_p(\mathbb{R}^n)}+\sup_{t\in[0,T]}\|f(t,\cdot)\|_{C^\theta_p(\mathbb{R}^n)}\right). \tag{1.2}$$

Note that for $p\equiv1$, Theorem 1.2 gives an optimal regularity result in the usual Hölder spaces. Owing to the lack of regularity of the coefficients with respect to the variable t, we do not expect the solution to problem (1.1) to be continuously t-differentiable in the whole of $[0,T]\times\mathbb{R}^n$ even if the coefficients are smooth with respect to x. Nevertheless, the discontinuity of the coefficients does not influence the regularity of the solution with respect to the space variables. This is not surprising since in [6] a similar result has been proved in bounded domains.

To solve problem (1.1), we first consider, in section 3, the particular case where Q, B, and c are independent of x. In such a case, we are able to find out an explicit representation of the solution u in terms of the data.

Then in section 4, we solve the Cauchy problem (1.1) using the classical method of continuity. First, in section 4.1, we find an a priori estimate for the solution to

problem (1.1). More precisely, we prove that if u solves the Cauchy problem (1.1) in the sense of Definition 1.1, then there exists a positive constant C such that

$$\sup_{t\in[0,T]}\|u(t,\cdot)\|_{C_p^{2+\theta}(\mathbb{R}^n)}\le C\bigg(\|u_0\|_{C_p^{2+\theta}(\mathbb{R}^n)}+\sup_{t\in[0,T]}\|f(t,\cdot)\|_{C_p^{\theta}(\mathbb{R}^n)}\bigg). \tag{1.3}$$

There are several methods to get (1.3). One consists of observing that the Krylov maximum principle holds in our situation and then applying a generalized version of the maximum principle in [5]. Here we prefer to apply the classical method of freezing of coefficients. For this purpose, we heavily use the boundedness of the coefficients in the whole of $[0,T]\times\mathbb{R}^n$

Then in section 4.2, thanks to the a priori estimate, we show, using a classical perturbation argument, that the Cauchy problem (1.1) admits a solution in the sense of Definition 1.1.

Finally, in the appendix, we prove four technical lemmas that have been used throughout the paper.

2. Notations and preliminaries.

DEFINITION 2.1. *For any $k\in\mathbb{N}$, $BUC^k(\mathbb{R}^n)$ denotes the Banach space of all the bounded and continuously differentiable up to the kth-order functions f for which $D^\alpha f$ is uniformly continuous in $\mathbb{R}^n$ for any α with length k. It is normed by*

$$\|f\|_{BUC^k(\mathbb{R}^n)}=\sum_{|\alpha|=0}^{k}\|D^\alpha f\|_\infty. \tag{2.1}$$

DEFINITION 2.2. *For any $T>0$, $\mathrm{Lip}\,([0,T]\times\mathbb{R}^n)$ denotes the vector space of all the functions $f:[0,T]\times\mathbb{R}^n\to\mathbb{R}$ such that*

$$\|f(t_2,x_2)-f(t_1,x_1)\|\le C\big(|t_2-t_1|^2+|x_2-x_1|^2\big)^{1/2}\quad\forall(t_j,x_j)\in[0,T]\times\mathbb{R}^n,\quad j=1,2, \tag{2.2}$$

for some positive constant C. Moreover, we define

$$[f]_{\mathrm{Lip}\,([0,T]\times\mathbb{R}^n)}=\inf\{C:\ (2.2)\ \textit{holds}\}$$

and we norm $\mathrm{Lip}\,([0,T]\times\mathbb{R}^n)$ by

$$\|f\|_{\mathrm{Lip}\,([0,T]\times\mathbb{R}^n)}=\|f\|_{L^\infty([0,T]\times\mathbb{R}^n)}+[f]_{\mathrm{Lip}\,([0,T]\times\mathbb{R}^n)}. \tag{2.3}$$

DEFINITION 2.3. *For any $\theta\in\mathbb{R}\backslash\mathbb{N}$, $C^\theta(\mathbb{R}^n)$ denotes the Banach space of all the functions f that are bounded and continuously differentiable up to the $[\theta]$-order such that for any $\alpha\in\mathbb{N}^n$ with length $[\theta]$,*

$$[D^\alpha f]_{\theta-[\theta]}:=\sup_{x,y\in\mathbb{R}^n,\ x\ne y}\frac{|D^\alpha f(y)-D^\alpha f(x)|}{|y-x|^{\theta-[\theta]}}<+\infty.$$

$C^\theta(\mathbb{R}^n)$ is normed by

$$\|f\|_{C^\theta(\mathbb{R}^n)}=\sum_{|\alpha|=0}^{[\theta]}\|D^\alpha f\|_\infty+\sum_{|\alpha|=[\theta]}[D^\alpha f]_{\theta-[\theta]}. \tag{2.4}$$

Throughout this paper, we will consider two families of weight functions: the polynomial weights, defined by $p(x)=1+|x|^{2m}$ for any $m\in\mathbb{N}\cup\{0\}$ and any $x\in\mathbb{R}^n$

and the exponential weights, defined by $p(x) = \exp\left((1 + \langle Sx, x\rangle)^{1/2}\right)$ for any $x \in \mathbb{R}^n$, S being any strictly positive and symmetric matrix. Moreover, we will deal with the following weighted spaces.

DEFINITION 2.4. *For any $k \in \mathbb{N}$ and any $\theta \in \mathbb{R} \cap \mathbb{N}^c$, $UC_p^k(\mathbb{R}^n)$ and $C_p^\theta(\mathbb{R}^n)$ denote the Banach spaces of all the functions f such that f/p belongs to $BUC^k(\mathbb{R}^n)$ and $C^\theta(\mathbb{R}^n)$, respectively. They are normed by*

$$\|f\|_{UC_p^k(\mathbb{R}^n)} = \|f/p\|_{BUC^k(\mathbb{R}^n)} \quad \text{and} \quad \|f\|_{C_p^\theta(\mathbb{R}^n)} = \|f/p\|_{C^\theta(\mathbb{R}^n)}. \tag{2.5}$$

In what follows, we will often use the following characterization of the previous weighted spaces.

LEMMA 2.5. *For any $k \in \mathbb{N}$ and any $\theta \in \mathbb{R}_+ \cap \mathbb{N}^c$, the following characterizations hold:*

(i) *$UC_p^k(\mathbb{R}^n)$ consists of all the continuously differentiable up to the kth-order functions f such that $(D^\alpha f)/p$ belongs to $BUC(\mathbb{R}^n)$ for any α with $|\alpha| \le k$. Moreover, the norm*

$$|\!|\!| f |\!|\!|_{UC_p^k(\mathbb{R}^n)} = \sum_{|\alpha|=0}^{k} \left\| \frac{D^\alpha f}{p} \right\|_\infty \qquad \forall f \in UC_p^k(\mathbb{R}^n)$$

is equivalent to the norm defined in (2.5).

(ii) *$C_p^\theta(\mathbb{R}^n)$ consists of all the continuously differentiable up to the $[\theta]$-order functions f such that for any $\alpha \in \mathbb{N}^n$ with $|\alpha| \le k$, $(D^\alpha f)/p$ belongs to $C^{\theta-[\theta]}(\mathbb{R}^n)$. Moreover, the norm*

$$|\!|\!| f |\!|\!|_{C_p^\theta(\mathbb{R}^n)} = \sum_{|\alpha|=0}^{[\theta]} \left\| \frac{D^\alpha f}{p} \right\|_\infty + \sum_{|\alpha|=[\theta]} \left[\frac{D^\alpha f}{p} \right]_{\theta-[\theta]} \qquad \forall f \in C_p^\theta(\mathbb{R}^n)$$

is equivalent to the norm defined in (2.5).

3. The case of coefficients not depending on x.

3.1. The case $f \equiv 0$. In this subsection, we are concerned with the following problem: *determine a function $u : [r, T] \times \mathbb{R}^n \to \mathbb{R}$ $(0 \le r < T)$ solution to the Cauchy problem*

$$\begin{cases} D_t u(t,x) = \displaystyle\sum_{i,j=1}^{n} q_{i,j}(t) D_{i,j} u(t,x) + \sum_{i=1}^{n} b_i(t) D_i u(t,x) + c(t) u(t,x), \\ \qquad\qquad (t,x) \in [r,T] \times \mathbb{R}^n, \\ u(r,x) = u_0(x), \qquad x \in \mathbb{R}^n, \end{cases} \tag{3.1}$$

under the following assumptions:

H1. $q_{i,j}(\cdot), b_j(\cdot)$ $(i,j = 1, \dots, n)$ and $c(\cdot)$ are measurable functions in $[0,T]$ bounded by a positive constant M;

H2. there exists a positive constant C_0 such that $\sum_{i,j=1}^n q_{i,j}(t)\xi_i\xi_j \ge C_0$ for any $|\xi| = 1$ almost everywhere (a.e.) in $[0,T]$.

We shall denote by $\mathcal{A}(t)$ the differential operator in the right-hand side of (3.1), by $Q_0(t)$ the matrix with elements $(Q_0)_{i,j}(t) = q_{i,j}(t)$, and by $B_0(t)$ the vector with components $(B_0)_j(t) = b_j(t)$.

By formally applying the Fourier transform to our Cauchy problem, we get the following representation for the solution to (3.1):

$$u(t,x)=\frac{\exp(C(t,r))}{(4\pi)^{n/2}(\det Q(t,r))^{1/2}}\int_{\mathbb{R}^n}\exp\left(-\frac{1}{4}\langle Q(t,r)^{-1}y,y\rangle\right)u_0(x-y+B(t,r))dy, \tag{3.2}$$

where

$$(Q(t,r))_{i,j}=\int_r^t q_{i,j}(s)ds,\quad (B(t,r))_j=\int_r^t b_j(s)ds,\quad C(t,r)=\int_r^t c(s)ds. \tag{3.3}$$

We now consider the family of linear operators $\{G(t,r)\}_{\{0\le r<t\}}$ defined by

$$\begin{aligned}G(t,r)\varphi(x)=&\frac{\exp(C(t,r))}{(4\pi)^{n/2}(\det Q(t,r))^{1/2}}\\&\times\int_{\mathbb{R}^n}\exp\left(-\frac{1}{4}\langle Q(t,r)^{-1}(y+B(t,r)),y+B(t,r)\rangle\right)\varphi(x-y)dy\end{aligned}$$

for any $\varphi\in UC_p(\mathbb{R}^n)$, p being either the polynomial or the exponential weight function.

LEMMA 3.1. *For any $\varphi\in UC_p(\mathbb{R}^n)$ and any $0\le r<s<t\le T$, $G(t,r)\varphi=G(t,s)G(s,r)\varphi$.*

Proof. We observe that for any $0\le r<s<t\le T$, $Q(t,r)=Q(t,s)+Q(s,r)$, $B(t,r)=B(t,s)+B(s,r)$, $C(t,r)=C(t,s)+C(s,r)$. Then, taking advantage of the Fubini–Tonelli theorem and integrating by parts, it is easy to prove the assertion. □

THEOREM 3.2. *Let $p:\mathbb{R}^n\to\mathbb{R}$ be defined by $p(x)=1+|x|^{2m}$ ($m\in\mathbb{N}\cup\{0\}$). Then for any $0\le r<t\le T$, $G(t,r)$ is a bounded linear operator mapping $UC_p(\mathbb{R}^n)$ into $UC_p^3(\mathbb{R}^n)$. Moreover, there exist positive constants $C(k,l,T)$ ($k,l=0,\dots,3$, $k\le l$) depending only on k, l, T, the sup norm of the coefficients, and on the constant C_0 in* H2 *such that*

$$\|G(t,r)\|_{\mathcal{L}(UC_p^k(\mathbb{R}^n);UC_p^l(\mathbb{R}^n))}\le C(k,l,T)(t-r)^{-(l-k)/2},\qquad 0\le l\le k\le 3. \tag{3.4}$$

Proof. We begin by showing that estimate (3.4) holds true when $k=l$. We observe that

$$\begin{aligned}\frac{p(x-y+B(t,r))}{p(x)}&=\frac{1+|x-y+B(t,r)|^{2m}}{1+|x|^{2m}}\\&\le 2^{2m-1}+2^{4m-2}\big(|y|^{2m}+n^mM^{2m}|t-r|^{2m}\big)\end{aligned} \tag{3.5}$$

for any $0\le r<t\le T$ and x, $y\in\mathbb{R}^n$. Moreover,

$$\begin{aligned}\left\|D_x\left(\frac{p(\cdot-y+B(t,r))}{p(\cdot)}\right)\right\|_\infty&\le 2^{2m-1}m\Big[3+2^{2m}\big(|y|^{2m}+n^mM^{2m}|t-r|^{2m}\big)\\&\quad+2^{2m-2}\big(|y|^{2m-1}+n^{m-1/2}M^{2m-1}|t-r|^{2m-1}\big)\Big]\end{aligned} \tag{3.6}$$

for any $y\in\mathbb{R}^n$ and any $0\le r<t$. Taking advantage of (3.5) and (3.6), it is easy to check that for any $\varphi\in UC_p(\mathbb{R}^n)$, the function $G(t,r)\varphi$ belongs to $UC_p(\mathbb{R}^n)$ and there exists a positive constant $C(0,T)$ such that

$$\|G(t,r)\varphi\|_{UC_p(\mathbb{R}^n)}\le C(0,T)\|\varphi\|_{UC_p(\mathbb{R}^n)},\qquad 0\le r<t\le T. \tag{3.7}$$

Then we observe that for any $\varphi \in UC_p^k(\mathbb{R}^n)$ $(k \le 3)$,

$$D^\alpha G(t,r)\varphi = G(t,r)D^\alpha \varphi \qquad \forall |\alpha| \le k. \tag{3.8}$$

Therefore, estimate (3.4) follows, in this particular case, from (3.7), (3.8), and Lemma 2.5.

We now consider the case $0 = k < l \le 3$ and observe that for any $\varphi \in UC_p(\mathbb{R}^n)$ and any $j = 1, \dots, n$, the function $G(t,r)\varphi$ is differentiable in $\mathbb{R}^n$ and its derivatives up to the third order are given by the following formulas:

$$\begin{aligned} D_j G(t,r)\varphi(x) = &-\frac{\exp(C(t,r))}{2(4\pi)^n}\int_{\mathbb{R}^n} (Q(t,r)^{-1/2}y)_j \exp\left(-\frac{1}{4}|y|^2\right) \\ &\times \varphi(x - Q(t,r)^{1/2}y + B(t,r))dy; \end{aligned} \tag{3.9}$$

$$\begin{aligned} D_i D_j G(t,r)\varphi(x) = &-\frac{\exp(C(t,r))}{2(4\pi)^n}\int_{\mathbb{R}^n} (Q(t,r)^{-1})_{j,i} \exp\left(-\frac{1}{4}|y|^2\right) \\ &\times \varphi(x - Q(t,r)^{1/2}y + B(t,r))dy \\ &+ \frac{\exp(C(t,r))}{4(4\pi)^n}\int_{\mathbb{R}^n} (Q(t,r)^{-1/2}y)_i (Q(t,r)^{-1/2}y)_j \\ &\times \exp\left(-\frac{1}{4}|y|^2\right)\varphi(x - Q(t,r)^{1/2}y + B(t,r))dy; \end{aligned} \tag{3.10}$$

$$\begin{aligned} D^3_{i,j,k} G(t,r)\varphi(x) = &\frac{\exp(C(t,r))}{4(4\pi)^n}\left[\int_{\mathbb{R}^n} (Q(t,r)^{-1})_{j,k}(Q(t,r)^{-1/2}y)_i \right. \\ &\times \exp\left(-\frac{1}{4}|y|^2\right)\varphi(x - Q(t,r)^{1/2}y + B(t,r))dy \\ &+ \int_{\mathbb{R}^n} (Q(t,r)^{-1})_{k,i}(Q(t,r)^{-1/2}y)_j \exp\left(-\frac{1}{4}|y|^2\right) \\ &\times \varphi(x - Q(t,r)^{1/2}y + B(t,r))dy \\ &+ \int_{\mathbb{R}^n} (Q(t,r)^{-1})_{j,i}(Q(t,r)^{-1/2}y)_k \exp\left(-\frac{1}{4}|y|^2\right) \\ &\times \varphi(x - Q(t,r)^{1/2}y + B(t,r))dy \\ &+ \frac{1}{2}\int_{\mathbb{R}^n} (Q(t,r)^{-1/2}y)_i (Q(t,r)^{-1/2}y)_j (Q(t,r)^{-1/2}y)_k \\ &\left. \times \exp\left(-\frac{1}{4}|y|^2\right)\varphi(x - Q(t,r)^{1/2}y + B(t,r))dy\right]. \end{aligned} \tag{3.11}$$

We recall here that $Q(t,r)^{1/2}$ is defined for any $0 \le r < t \le T$ by the following formula:

$$Q(t,r)^{1/2} = \frac{1}{2\pi \mathrm{i}}\int_\gamma \sqrt{\lambda}(\lambda I - Q(t,r))^{-1}d\lambda, \tag{3.12}$$

where γ is any oriented envelope of the spectrum of $Q(t,r)$ contained in the angle $\Sigma = \{\lambda \in \mathbb{C} : \arg\lambda \in [0, \pi/2]\}$ and $\sqrt{\lambda}$ is the principal value of the square root in $\mathbb{C}$.

Then, using estimates (3.5) and (3.6), we easily deduce that for any multiindex α with $|\alpha| \le 3$, the functions $D^\alpha G(t,r)\varphi \in UC_p(\mathbb{R}^n)$. Moreover, there exist positive

constants $\widetilde{C}_k(T)$ $(k=1,2,3)$ such that

$$\|D_jG(t,r)\varphi\|_{UC_p(\mathbb{R}^n)}\le\widetilde{C}_1(T)(t-r)^{-1/2}\|\varphi\|_{UC_p(\mathbb{R}^n)},\qquad 0\le r<t\le T; \tag{3.13}$$

$$\|D^2_{i,j}G(t,r)\varphi\|_{UC_p(\mathbb{R}^n)}\le\widetilde{C}_2(T)(t-r)^{-1}\|\varphi\|_{UC_p(\mathbb{R}^n)},\qquad 0\le r<t\le T; \tag{3.14}$$

$$\|D^3_{i,j,k}G(t,r)\varphi\|_{UC_p(\mathbb{R}^n)}\le\widetilde{C}_3(T)(t-r)^{-3/2}\|\varphi\|_{UC_p(\mathbb{R}^n)},\qquad 0\le r<t\le T. \tag{3.15}$$

From (3.13)–(3.15) we deduce that there exist three positive constants $C(k,T)$ $(k=1,2,3)$ depending only on the sup norm of the coefficients and on the constant C_0 in H2 such that (3.4) holds.

To conclude, we consider the case $0<k<l\le 3$ and observe that for any $\alpha\in\mathbb{N}^n$ such that $|\alpha|>k$, there exists $\beta\in\mathbb{N}^n$ such that $|\beta|=k$ and

$$D^\alpha G(t,r)\varphi=D^{\alpha-\beta}D^\beta G(t,r)\varphi=D^{\alpha-\beta}G(t,r)D^\beta\varphi.$$

Then (3.4) follows from the previous two cases. □

We now consider the case of exponential weight function.

THEOREM 3.3. *Let $p:\mathbb{R}^n\to\mathbb{R}$ be defined by $p(x)=\exp((1+\langle Sx,x\rangle)^{1/2})$. Then for any $0\le r<t\le T$, $G(t,r)$ is a bounded linear operator mapping $UC_p(\mathbb{R}^n)$ into $UC_p^3(\mathbb{R}^n)$. Moreover, for any k, $l=0,\dots,3$, $k\le l$, there exists a positive constant $D(k,l,T)$ depending on the sup norm of the coefficients, on T, and on the constant C_0 in* H2 *such that*

$$\|G(t,r)\|_{\mathcal{L}(UC_p^k(\mathbb{R}^n);UC_p^l(\mathbb{R}^n))}\le D(k,l,T)(t-r)^{-(l-k)/2},\qquad 0\le k\le l\le 3. \tag{3.16}$$

Proof. The proof is similar to the one given in the case of the polynomial weight function. In fact, we observe that

$$\begin{aligned}\frac{p(x-y+B(t,r))}{p(x)}&=\exp\Big((1+|S^{1/2}(x-y+B(t,r))|^2)^{1/2}-(1+|S^{1/2}x|^2)^{1/2}\Big)\\&\le\exp\Big(|S^{1/2}(y-B(t,r))|\Big)\\&\le\exp\Big(\|S^{1/2}\|Mn^{1/2}(t-r)\Big)\exp\big(\|S^{1/2}\||y|\big)\end{aligned} \tag{3.17}$$

and

$$\left\|D_x\left(\frac{p(\cdot-y+B(t,r))}{p(\cdot)}\right)\right\|_\infty\le 2\|S^{1/2}\|\exp\Big(\|S^{1/2}\|Mn^{1/2}(t-r)\Big)\exp\big(\|S^{1/2}\||y|\big) \tag{3.18}$$

for any x, $y\in\mathbb{R}^n$, $0\le r<t\le T$. To derive estimate (3.18), it suffices to observe that

$$\begin{aligned}\left|D_x\left(\frac{p(x-y+B(t,r))}{p(x)}\right)\right|&=\left|\frac{S(x-y+B(t,r))}{\big(1+\|S^{1/2}(x-y+B(t,r))\|^2\big)^{1/2}}-\frac{Sx}{(1+\|S^{1/2}x\|^2)^{1/2}}\right|\\&\quad\times\frac{p(x-y+B(t,r))}{p(x)}\le 2\|S^{1/2}\|\frac{p(x-y+B(t,r))}{p(x)}.\end{aligned}$$

Therefore, reasoning as in the case of the polynomial weight function, it can be easily proved that for any $\varphi\in UC_p^k(\mathbb{R}^n)$, $G(t,r)\varphi\in UC_p^k(\mathbb{R}^n)$ $(k=0,\dots,3)$ and fulfills the following estimate (cf. Lemma 2.5):

$$|\!|\!|G(t,r)\varphi|\!|\!|_{UC_p^k(\mathbb{R}^n)}\le\widetilde{d}_0\exp\Big((M+Mn^{1/2}\|S^{1/2}\|)T\Big)|\!|\!|\varphi|\!|\!|_{UC_p^k(\mathbb{R}^n)} \tag{3.19}$$

for any $0 \le r < t \le T$ and any $k = 0, \dots, 3$, where

$$\widetilde{d}_j = (4\pi)^{-n/2} \int_{\mathbb{R}^n} |y|^j \exp(n^{1/2} M^{1/2} \|S^{1/2}\| |y|) \exp\Big(-\frac{1}{4}|y|^2 \Big) dy, \qquad j = 0, \dots, 3.$$

Therefore, thanks to Lemma 2.5, we get (3.16) in the case $k = l$.

Next we observe that for any $\varphi \in UC_p(\mathbb{R}^n)$, $G(t,r)\varphi \in UC_p^3(\mathbb{R}^n)$, and (cf. (3.9)–(3.11))

$$\|D_i G(t,r)\varphi\|_{UC_p(\mathbb{R}^n)} \le \frac{C_0^{-1/2} \widetilde{d}_1}{2(t-r)^{1/2}} \exp\Big(M(1 + n^{1/2}\|S^{1/2}\|)T \Big) \|\varphi\|_{UC_p(\mathbb{R}^n)}; \tag{3.20}$$

$$\|D_{i,j}^2 G(t,r)\varphi\|_{UC_p(\mathbb{R}^n)} \le \frac{C_0^{-1}(2\widetilde{d}_0 + \widetilde{d}_2)}{4(t-r)} \exp\Big(M(1 + n^{1/2}\|S^{1/2}\|)T \Big) \|\varphi\|_{UC_p(\mathbb{R}^n)}; \tag{3.21}$$

$$\|D_{i,j,k}^3 G(t,r)\varphi\|_{UC_p(\mathbb{R}^n)} \le \frac{C_0^{-3/2}(6\widetilde{d}_1 + \widetilde{d}_3)}{8(t-r)^{3/2}} \exp\Big(M(1 + n^{1/2}\|S^{1/2}\|)T \Big) \|\varphi\|_{UC_p(\mathbb{R}^n)} \tag{3.22}$$

for any $i, j, k = 1, \dots, n$ and any $0 \le r < t \le T$. Then, taking Lemma 2.5 into account, from (3.20)–(3.22) we easily deduce (3.16). The case $0 < k < l \le 3$ can be proved as in Theorem 3.2. □

The estimates of Theorems 3.2 and 3.3 may be extended by interpolation to weighted Hölder norms, thanks to the following lemma.

LEMMA 3.4. *Let p be either the polynomial or the exponential weight function. Then for any $\sigma \in (0,1)$, any $0 \le \alpha < \beta$, and any $m \in \mathbb{N}$ such that $\alpha + \sigma(\beta - \alpha) \notin \mathbb{N}$ and $\sigma m \notin \mathbb{N}$, it holds that*

$$(UC_p(\mathbb{R}^n); UC_p^m(\mathbb{R}^n))_{\sigma,\infty} = C_p^{\sigma m}(\mathbb{R}^n);$$
$$(UC_p^\alpha(\mathbb{R}^n); UC_p^\beta(\mathbb{R}^n))_{\sigma,\infty} = C_p^{\alpha + \sigma(\beta-\alpha)}(\mathbb{R}^n),$$

with equivalence of the norms.

Proof. See [11, Theorems 1.2, 1.3, 1.7, and 1.8]. □

THEOREM 3.5. *Let p be the polynomial or the exponential weight function. Then for any $0 \le \alpha \le \beta \le 3$ and any $0 \le r < t$, $G(t,r)$ is a bounded linear operator mapping $UC_p^\alpha(\mathbb{R}^n)$ into $UC_p^\beta(\mathbb{R}^n)$. Moreover, there exists a positive constant $C(\alpha,\beta)$ such that for any $0 \le r < t \le T$,*

$$\|G(t,r)\|_{\mathcal{L}(UC_p^\alpha(\mathbb{R}^n); UC_p^\beta(\mathbb{R}^n))} \le C(\alpha,\beta)(t-r)^{-(\beta-\alpha)/2}. \tag{3.23}$$

Proof. It is sufficient to use the interpolation arguments of [11, Theorems 1.4 and 1.9]. □

The regularity properties of $G(t,r)\varphi$ are proved by the next lemmas. First, we deal with strong continuity.

LEMMA 3.6. *Suppose that p is either the polynomial or the exponential weight function. Then for any $\varphi \in UC_p(\mathbb{R}^n)$, $G(t,r)\varphi$ tends to φ in $UC_p(\mathbb{R}^n)$ as t tends to r.*

Proof. We start considering the case of polynomial weight function and observe that

$$\begin{aligned}
&\left|\frac{p(x-y+B(t,r))}{p(x)}-1\right| \\
&\quad=\left|\frac{|x-y+B(t,r)|^{2m}-|x|^{2m}}{1+|x|^{2m}}\right| \\
&\quad\le(1+|x|)^{-2m}\sum_{j=1}^{m}\binom{m}{k}|x|^{2(m-k)}\left||y-B(t,r)|^2-2\langle x,y-B(t,r)\rangle\right|^k \\
&\quad\le\sum_{k=1}^{m}\sum_{j=0}^{k}\binom{m}{k}\binom{k}{j}2^{2k-1}\Big(|y|^{2k-j}+\|B(t,r)\|^{2k-j}\Big). \qquad (3.24)
\end{aligned}$$

Then, taking advantage of (3.5) and (3.24) and setting

$$d_l=(4\pi)^{-n/2}\int_{\mathbb{R}^n}|y|^l\exp\left(-\frac{1}{4}|y|^2\right)dy,\qquad l\in\mathbb{N}\cup\{0\},$$

we deduce that for any $\varphi\in UC_p^1(\mathbb{R}^n)$,

$$\begin{aligned}
&\left|\frac{G(t,r)\varphi(x)}{1+|x|^{2m}}-\frac{\varphi(x)}{1+|x|^{2m}}\right| \\
&=\frac{\exp(C(t,r))}{(4\pi)^{n/2}}\int_{\mathbb{R}^n}\exp\left(-\frac{1}{4}|y|^2\right)\left|\frac{\varphi(x-Q(t,r)^{1/2}y+B(t,r))}{p(x-Q(t,r)^{1/2}y+B(t,r))}-\frac{\varphi(x)}{p(x)}\right| \\
&\quad\times\frac{p(x-Q(t,r)^{1/2}y+B(t,r))}{p(x)}dy \\
&\quad+\frac{\exp(C(t,r))}{(4\pi)^{n/2}}\int_{\mathbb{R}^n}\exp\left(-\frac{1}{4}|y|^2\right)\left|\frac{\varphi(x)}{p(x)}\right|\left|\frac{p(x-Q(t,r)^{1/2}y+B(t,r))}{p(x)}-1\right|dy \\
&\quad+|\exp(C(t,r))-1|\|\varphi\|_{UC_p(\mathbb{R}^n)} \\
&\le\exp(C(t,r))\|D\varphi\|_{UC_p(\mathbb{R}^n)}\Big[2^{2m-1}d_1(nM)^{1/2}|t-r|^{1/2}\Big(1+2^{2m-1}n^mM^{2m}|t-r|^{2m}\Big) \\
&\quad+2^{4m-2}d_{2m}n^{m+1/2}M^{m+1}|t-r|^{m+1}+2^{4m-2}d_{2m+1}n^{m+1/2}M^{m+1/2}|t-r|^{m+1/2} \\
&\quad+2^{2m-1}n^{1/2}M|t-r|+2^{4m-2}M^{2m+1}n^{m+1/2}|t-r|^{2m+1}\Big] \\
&\quad+|\exp(C(t,r))-1|\|\varphi\|_{UC_p(\mathbb{R}^n)} \\
&\quad\times\sum_{k=1}^{m}\sum_{j=0}^{k}\binom{m}{k}\binom{k}{j}2^{2k-1}\Big(d_{2k+j}(nM)^{k-j/2}|t-r|^{k-j/2}+n^{k-j/2}M^{2k-j}|t-r|^{2k-j}\Big),
\end{aligned}$$

and the right side of the previous inequality tends to 0 as t tends to r. Then by density, taking (3.4) into account, we can prove that for any $\varphi\in UC_p(\mathbb{R}^n)$, $G(t,r)\varphi$ tends uniformly to φ as t tends to r.

We now move on to consider the case of exponential weight function. We observe that for any $y\in\mathbb{R}^n$ and $0\le r<t$ we have

$$\begin{aligned}
\left\|\frac{p(\cdot-Q(t,r)^{1/2}y+B(t,r))}{p(\cdot)}-1\right\|_\infty&\le\exp\big(\|S^{1/2}\||B(t,r)|+\|S^{1/2}\||y|\big) \\
&\quad-\exp\big(-\|S^{1/2}\||B(t,r)|-\|S^{1/2}\||y|\big)
\end{aligned}$$

$$\le 2\|S^{1/2}\|(nM)^{1/2}\Big(|t-r|^{1/2}|y| + M^{1/2}|t-r|\Big)$$
$$\times \exp\Big(\|S^{1/2}\|(nMT)^{1/2}\big(|y| + (nMT)^{1/2}\big)\Big). \tag{3.25}$$

Using estimates (3.17) and (3.25) and reasoning as in the case of polynomial weight function, the assertion can be proved. □

In the next lemma we describe the smoothing properties of $G(t,r)$.

LEMMA 3.7. *Suppose that p is either the polynomial or the exponential weight function. Then for any $\varphi \in UC_p(\mathbb{R}^n)$ and any $r \in [0,T)$, the function $u : [r,T]\times\mathbb{R}^n \to \mathbb{R}$ defined by $u(t,x) = G(t,r)\varphi(x)$ is twice differentiable with respect to the variable x in $[0,T]\times\mathbb{R}^n$ and it is differentiable with respect to the variable t for any $t \in E_r \times \mathbb{R}^n$, where E_r is defined by*

$$E_r = \{t \in (r,T] :\ Q,\ B,\ C \text{ are differentiable at } t\},$$

and Q, B, and C are defined in (3.3). *Moreover, $D_t u(t,x) = \mathcal{A}(t)u(t,x)$ for any $(t,x) \in E_r \times \mathbb{R}^n$.*

Proof. We begin the proof by considering $\varphi \in UC_p^2(\mathbb{R}^n)$ and observing that E_r is a measurable set and the measure of $[r,T]\backslash E_r$ is equal to zero. Moreover, from (3.12) we deduce that $Q(t,r)^{1/2}$ is differentiable with respect to the variable t at any $t \in E_r$. Hence we can easily show that u is differentiable with respect to the variable t at each point $(t,x) \in E_r \times \mathbb{R}^n$ and

$$D_t u(t,x) = c(t)u(t,x) + (4\pi)^{-n/2}\exp(C(t,r))$$
$$\times \int_{\mathbb{R}^n} \left\langle (D\varphi)(x + Q(t,r)^{1/2}y + B(t,r)), \frac{d}{dt}(Q(t,r)^{1/2})y \right\rangle \exp\left(-\frac{1}{4}|y|^2\right) dy$$
$$+ (4\pi)^{-n/2}\exp(C(t,r)) \int_{\mathbb{R}^n} \langle (D\varphi)(x + Q(t,r)^{1/2}y + B(t,r)), B_0(t)\rangle \exp\left(-\frac{1}{4}|y|^2\right) dy.$$

Then, integrating by parts the first integral that occurs in the definition of $D_t u$, we get

$$\int_{\mathbb{R}^n} \left\langle (D\varphi)(x + Q(t,r)^{1/2}y + B(t,r)), \frac{d}{dt}(Q(t,r)^{1/2})y \right\rangle \exp\left(-\frac{1}{4}|y|^2\right) dy$$
$$= \sum_{i,j=1}^n \left(\frac{d}{dt}(Q(t,r)^{1/2})\right)_{i,j} \int_{\mathbb{R}^n} y_j \exp\left(-\frac{1}{4}|y|^2\right)(D_i\varphi)(x + Q(t,r)^{1/2}y + B(t,r))dy$$
$$= 2\sum_{i,j,k=1}^n \left(\frac{d}{dt}(Q(t,r)^{1/2})\right)_{i,j} (Q(t,r)^{1/2})_{k,j}$$
$$\times \int_{\mathbb{R}^n} (D_{i,k}\varphi)(x + Q(t,r)^{1/2}y + B(t,r)) \exp\left(-\frac{1}{4}|y|^2\right) dy$$
$$= \sum_{i,k=1}^n \left(\frac{d}{dt}(Q(t,r)^{1/2})Q(t,r)^{1/2}\right)_{i,k}$$
$$\times \int_{\mathbb{R}^n} (D_{i,k}\varphi)(x + Q(t,r)^{1/2}y + B(t,r)) \exp\left(-\frac{1}{4}|y|^2\right) dy$$
$$+ \sum_{i,k=1}^n \left(\frac{d}{dt}(Q(t,r)^{1/2})Q(t,r)^{1/2}\right)_{k,i}$$

$$\times \int_{\mathbb{R}^n} (D_{i,k}\varphi)(x+Q(t,r)^{1/2}y+B(t,r))\exp\left(-\frac{1}{4}|y|^2\right)dy$$
$$= \int_{\mathbb{R}^n} \mathrm{Tr}\left[\left(\frac{d}{dt}Q(t,r)^{1/2}Q(t,r)^{1/2}+Q(t,r)^{1/2}\frac{d}{dt}Q(t,r)^{1/2}\right)\right.$$
$$\left.\times\,(D^2\varphi)(x+Q(t,r)^{1/2}y+B(t,r))\right]\exp\left(-\frac{1}{4}|y|^2\right)dy$$
$$= \int_{\mathbb{R}^n} \mathrm{Tr}\big[Q_0(t)(D^2\varphi)(x+Q(t,r)^{1/2}y+B(t,r))\big]\exp\left(-\frac{1}{4}|y|^2\right)dy,$$

so that

$$D_t u(t,x) = \frac{\exp(C(t,r))}{(4\pi)^{n/2}}\int_{\mathbb{R}^n}\mathcal{A}(t)\varphi(x+Q(t,r)^{1/2}y+B(t,r))\exp\left(-\frac{1}{4}|y|^2\right)dy.$$

Then an elementary computation shows that $G(t,r)\mathcal{A}(t)\varphi = \mathcal{A}(t)(G(t,r)\varphi)$. Next we suppose that $\varphi \in UC_p(\mathbb{R}^n)$ and observe that there exists a sequence $\{\varphi_k\}_{k\in\mathbb{N}}$ belonging to $UC_p^2(\mathbb{R}^n)$ such that $\varphi_k \to \varphi$ in $UC_p(\mathbb{R}^n)$ as $k\to+\infty$. Taking advantage of (3.5) and (3.17), we deduce that for any $t\in(r,T]$, $G(t,r)\varphi_k$ tends to $G(t,r)\varphi$ as $k\to+\infty$ in $UC_p^2(\mathbb{R}^n)$, so that $\mathcal{A}(t)G(t,r)\varphi_k \to \mathcal{A}(t)G(t,r)\varphi$ for any $(t,x)\in(0,T]\times\mathbb{R}^n$. We now observe that $G(t,r)\varphi$ is differentiable with respect to the variable t at any point $(t,x)\in E_r\times\mathbb{R}^n$ and

$$(D_tG(t,r)\varphi)(x) = (4\pi)^{-n/2}D_t\Big(\exp(C(t,r))\big[\det Q(t,r)\big]^{-1/2}\Big)$$
$$\times\int_{\mathbb{R}^n}\exp\left(-\frac{1}{4}\langle Q(t,r)^{-1}(y+B(t,r)),y+B(t,r)\rangle\right)\varphi(x-y)dy$$
$$-\frac{\exp(C(t,r))}{4(4\pi)^{n/2}\det Q(t,r)^{1/2}}\int_{\mathbb{R}^n}\exp\left(-\frac{1}{4}\langle Q(t,r)^{-1}\big(y+B(t,r)\big),y+B(t,r)\rangle\right)$$
$$\times\,\varphi(x-y)D_t\big(\langle Q(t,r)^{-1}\big(y+B(t,r)\big),y+B(t,r)\rangle\big)dy.$$

Therefore, $G(t,r)\varphi$ is differentiable with respect to the variable t in $(r,T]\times\mathbb{R}^n$. Moreover, $D_t\varphi_k(t,x)$ tends to $D_t\varphi(t,x)$ as $k\to+\infty$ for any $(t,x)\in E_r\times\mathbb{R}^n$ and the proof is now complete. □

3.2. The nonhomogeneous case. In this subsection, we will consider the Cauchy problem

$$\text{(3.26)}\qquad \begin{cases} u_t(t,x) = \mathcal{A}(t)u(t,x)+f(t,x), & (t,x)\in[0,T]\times\mathbb{R}^n,\\ u(0,x) = u_0(x), & x\in\mathbb{R}^n. \end{cases}$$

under assumptions H1 and H2 (cf. section 3.1). We will denote by p either the polynomial or the exponential weight function.

We give the following definition of the "mild solution" to problem (3.26).

DEFINITION 3.8. *Suppose that*

(i) $f : [0,T] \times \mathbb{R}^n \to \mathbb{R}$ *is such that* $f/p \in L^\infty([0,T] \times \mathbb{R}^n)$ *and* $x \to f(r,x)$ *is a measurable function for any* $r \in [0,T]$;

(ii) $u_0 : \mathbb{R}^n \to \mathbb{R}$ *is such that* $u_0/p \in L^\infty(\mathbb{R}^n)$.

Then the function $u : [0,T] \times \mathbb{R}^n \to \mathbb{R}$ *defined by*

$$u(t,x) = G(t,0)u_0(x) + \int_0^t G(t,r)f(r,\cdot)(x)dr \tag{3.27}$$

is called the "mild" solution to problem (3.26).

In what follows, we deal with the space $B([0,T]; C_p^\theta(\mathbb{R}^n))$, $\theta \in (0,1)$.

DEFINITION 3.9. *For any* $T > 0$ *and any* $\theta \in \mathbb{R}_+$, $B([0,T]; C_p^\theta(\mathbb{R}^n))$ *denotes the vector space of all the functions* f *such that for any* $t \in [0,T]$, $f(t,\cdot) \in C_p^\theta(\mathbb{R}^n)$ *and* $\sup_{t\in[0,T]} \|f(t,\cdot)\|_{C_p^\theta(\mathbb{R}^n)} < +\infty$.

We are going to study the main properties of the "mild" solution to problem (3.26). The following lemma will be useful.

LEMMA 3.10. *Suppose that* $R \in L^\infty([0,T]; \mathbb{R}^{n^2})$, $S \in L^\infty([0,T]; \mathbb{R}^n)$, *and* $d \in L^\infty([0,T])$. *Then there exist three sequences* $\{R^{(k)}\}_{k\in\mathbb{N}} \in C([0,T]; \mathbb{R}^{n^2})$, $\{S^{(k)}\}_{k\in\mathbb{N}} \in C([0,T]; \mathbb{R}^n)$, *and* $\{d^{(k)}\}_{k\in\mathbb{N}} \in C([0,T])$ *such that*

(i) $R^{(k)}(t) \to R(t)$, $S^{(k)}(t) \to S(t)$, $d^{(k)}(t) \to d(t)$ *a.e. in* $[0,T]$ *as* $k \to +\infty$;

(ii) $\|R^{(k)}\|_{C([0,T];\mathbb{R}^{n^2})} \le \|R\|_{B([0,T];\mathbb{R}^{n^2})}$;

(iii) $\|S^{(k)}\|_{C([0,T];\mathbb{R}^n)} \le \|S\|_{B([0,T];\mathbb{R}^n)}$;

(iv) $\|d^{(k)}\|_{C([0,T])} \le \|d\|_{B([0,T])}$.

Moreover, if R *is uniformly strictly definite positive a.e. in* $[0,T]$, *then* $R^{(k)}$ *is also, independently of* k.

Proof. Let us consider the case of R, the others being similar. We define the matrix $R^{(k)}$ as follows:

$$(R^{(k)})_{i,j}(t) = \left(\frac{k}{4\pi}\right)^{1/2} \int_0^T r_{i,j}(s) \exp\left(-\frac{k}{4}|t-s|^2\right) ds, \qquad i,j = 1,\dots,n.$$

Then it is easy to check that $R^{(k)}$ admits a subsequence converging a.e. to the matrix R. Moreover, it can be easily proved that

$$\langle R^{(k)}x, x\rangle \ge (4\pi)^{-1/2} C \exp\left(-\frac{1}{4}T^2\right), \qquad x \in \mathbb{R}^n \qquad \forall k \in \mathbb{N}$$

and C is the constant of coercivity of R. □

LEMMA 3.11. *Let* θ, α *be two positive real numbers such that* $0 < \theta < \alpha < 1$. *Then for any interval* $I \subset \mathbb{R}$ *and any* $\varphi : I \to C_p^\theta(\mathbb{R}^n)$ *such that for any* $x \in \mathbb{R}^n$, *the real function* $t \to \varphi(t)(x)$ *is measurable in* I *and* $\|\varphi(t)\|_{C_p^\alpha(\mathbb{R}^n)} \le c(t)$ *(resp.,* $\|\varphi\|_{C_p^{2+\alpha}(\mathbb{R}^n)} \le c(t)$*) with* $c \in L^1([0,T])$, *the function*

$$f(x) = \int_I \varphi(t)(x)dt, \qquad x \in \mathbb{R}^n,$$

belongs to $C_p^\alpha(\mathbb{R}^n)$ *(resp.,* $C_p^{2+\alpha}(\mathbb{R}^n)$*) and there exists a positive constant* K, *independent of* φ, *such that*

$$\|f\|_{C_p^\alpha(\mathbb{R}^n)} \le K\|c\|_{L^1([0,T])} \qquad (\textit{resp.,}\ \|f\|_{C_p^{2+\alpha}(\mathbb{R}^n)} \le K\|c\|_{L^1([0,T])}).$$

Proof. See [12, section 3] and [11, Lemmas 3.1 and 3.2]. □

Thanks to Lemma 3.11 and Theorem 3.5, we can prove the following theorem. For a similar statement, see also [13, Theorem 2.2].

THEOREM 3.12. *Suppose that f is a measurable function belonging to $B([0,T];C_p^\theta(\mathbb{R}^n))$, $\theta\in(0,1)$. Then the function $v:[0,T]\times\mathbb{R}^n\to\mathbb{R}$ defined by*

$$v(t,x)=\int_0^t G(t,r)f(r,\cdot)(x)dr\qquad\forall(t,x)\in[0,T]\times\mathbb{R}^n$$

belongs to $B([0,T];C_p^{2+\theta}(\mathbb{R}^n))$ and there exists a positive constant C, independent of f, such that

$$\sup_{t\in[0,T]}\|v(t,\cdot)\|_{C_p^{2+\theta}(\mathbb{R}^n)}\le C\sup_{t\in[0,T]}\|f(t,\cdot)\|_{C_p^\theta(\mathbb{R}^n)}. \tag{3.28}$$

Proof. We begin the proof by remarking that $r\to G(t,r)f(r,\cdot)(x)$ is measurable in $[0,t]$ for any $(t,x)\in(0,T]\times\mathbb{R}^n$ (cf. the appendix, Lemma A.1). Next we recall that for any pair of Banach spaces X and Y with $Y\subset X$ and any $\beta\in(0,1)$, $(X,Y)_{\beta,\infty}$ denotes the vector space of all $x\in X$ such that $\sup_{t>0}t^{-\beta}K(t,x)<+\infty$, where $K(t,x)=\inf_{a+b=x,\ a\in X,\ b\in Y}\left(\|a\|_X+t\|b\|_Y\right)$ (see [15, Chapter 1]).

Then, following [13], we split $v(t)$ as $v(t)=a(\xi,t)+b(\xi,t)$, where

$$a(\xi,t)=\begin{cases}\displaystyle\int_0^\xi G(t,t-r)f(t-r,\cdot)(x)dr & \text{if } \xi\le t,\\ \displaystyle\int_0^t G(t,t-r)f(t-r,\cdot)(x)dr & \text{if } \xi>t;\end{cases}$$

$$b(\xi,t)=\begin{cases}\displaystyle\int_\xi^t G(t,t-r)f(t-r,\cdot)(x)dr & \text{if } \xi\le t,\\ 0 & \text{if } \xi>t.\end{cases}$$

Taking advantage of Lemma 3.11 and estimate (3.23), we deduce that $a(\xi,t)\in C_p^\alpha(\mathbb{R}^n)$ and $b(\xi,t)\in C_p^{2+\alpha}(\mathbb{R}^n)$ for every t and there exist two positive constants $C(\alpha,\theta)$, $C(2+\alpha,\theta)$ such that

$$\|a(\xi)\|_{C_p^\alpha(\mathbb{R}^n)}\le C(\alpha,\theta)\xi^{1-(\alpha-\theta)/2}\sup_{r\in[0,T]}\|f(r,\cdot)\|_{C_p^\theta(\mathbb{R}^n)};$$
$$\|b(\xi)\|_{C_p^{2+\alpha}(\mathbb{R}^n)}\le C(2+\alpha,\theta)\xi^{-(\alpha-\theta)/2}\sup_{r\in[0,T]}\|f(r,\cdot)\|_{C_p^\theta(\mathbb{R}^n)}.$$

Therefore, there exists a positive constant $\widetilde{C}(\alpha,\theta)$ such that

$$\xi^{-1+(\alpha-\theta)/2}K(\xi,v(t))\le\widetilde{C}(\alpha,\theta)\sup_{0\le r\le T}\|f(r,\cdot)\|_{C_p^\theta(\mathbb{R}^n)},\qquad\xi>0.$$

Hence we deduce that $v(t)\in(C_p^\alpha(\mathbb{R}^n),C_p^{2+\alpha}(\mathbb{R}^n))_{1-(\alpha-\theta)/2,\infty}$ and the statement follows from Lemma 3.4. □

LEMMA 3.13. *For any measurable function f belonging to $B([0,T];C_p^\theta(\mathbb{R}^n))$ and any $u_0\in C_p^{2+\theta}(\mathbb{R}^n)$, the function u defined by* (3.27) *is such that u/p is a Lipschitz continuous function in $[0,T]\times\mathbb{R}^n$.*

Proof. We begin by proving that for any $x\in\mathbb{R}^n$, $t\to u(t,x)$ is a Lipschitz continuous function in $[0,T]$. Let us show that the function $t\to G(t,0)u_0(x)$ is

Lipschitz continuous in $[0,T]$ for any $x \in \mathbb{R}^n$. For this purpose we consider the function $G_k(t,0)u_0$ defined by

$$G_k(t,0)u_0(x) = \frac{\exp\left(C^{(k)}(t,0)\right)}{(4\pi)^{n/2}} \times \int_{\mathbb{R}^n} \exp\left(-\frac{1}{4}|y|^2\right) u_0(x - Q^{(k)}(t,0)^{1/2}y + B^{(k)}(t,0))dy,$$

where $Q^{(k)}$, $B^{(k)}$, and $C^{(k)}$ are defined as in (3.3) with $q_{i,j}$, b_j, and c replaced by $q_{i,j}^{(k)}$, $b_j^{(k)}$, and $c^{(k)}$ defined in Lemma 3.10. From Lemma 3.7, we deduce that the function $(t,x) \to G_k(t,0)u_0(x)$ is differentiable in $[0,T]\times\mathbb{R}^n$ with respect to the variable t and solves in that strip the Cauchy problem (3.26) with $f \equiv 0$ and $q_{i,j}$, b_j, and c replaced by $q_{i,j}^{(k)}$, $b_j^{(k)}$, and $c^{(k)}$, respectively $(i,j = 1,\dots,n)$. Taking Lemma 3.10 and estimates (3.5), (3.17) into account, it is easy to show that $G_k(t,0)u_0(x)$ tends to $G(t,0)u_0(x)$ as $k \to +\infty$ for any $(t,x) \in [0,T]\times\mathbb{R}^n$. Moreover, from the same estimates quoted above, we deduce that there exists a positive constant $C_1(T)$, independent of x and $k \in \mathbb{N}$, such that

$$|D_x^j G_k(t,0)u_0(x)| \le C_1(T)p(x)\|u_0\|_{C_p^{2+\theta}(\mathbb{R}^n)} \tag{3.29}$$

for any $(t,x) \in [0,T]\times\mathbb{R}^n$, any $k \in \mathbb{N}$, and any $j = 0,1,2$ (cf. Theorems 3.2 and 3.3). Hence

$$|D_t G_k(t,0)u_0(x)| \le (n^2+n+1)MC_1(T)p(x)\|u_0\|_{C_p^{2+\theta}(\mathbb{R}^n)}$$

for any $(t,x) \in [0,T]\times\mathbb{R}^n$ and any $k \in \mathbb{N}$. Consequently, for any t_1, $t_2 \in [0,T]$,

$$|G_k(t_2,0)u_0(x) - G_k(t_1,0)u_0(x)| \le (n^2+n+1)MC_1(T)p(x)\|u_0\|_{C_p^{2+\theta}(\mathbb{R}^n)}|t_2 - t_1|.$$

As $k \to +\infty$, we deduce that $t \to G(t,0)u_0(x)$ is a Lipschitz continuous function and

$$[G(\cdot,0)u_0(x)]_{\mathrm{Lip}([0,T])} \le (n^2+n+1)MC_1(T)p(x) \qquad \forall x \in \mathbb{R}^n. \tag{3.30}$$

Let us consider the function

$$v(t,x) = \int_0^t G(t,r)f(r,\cdot)(x)dr.$$

Define the approximate semigroup $G_k(t,r)$ as in the case $r = 0$ and the function v_k by the formula

$$v_k(t,x) = \int_0^t G_k(t,r)f(r,\cdot)(x)dr.$$

As is easily seen, $v_k(t,x)$ tends to $v(t,x)$ as $k \to +\infty$ for any $(t,x) \in [0,T]\times\mathbb{R}^n$. From Lemma 3.7, we deduce that $t \to G_k(t,r)f(r,\cdot)(x)$ is differentiable with respect to the variable t in $[r,T]\times\mathbb{R}^n$. Moreover, there exists a positive constant $C_2(T)$, independent of k, such that

$$|D_x^j G_k(t,r)f(s,\cdot)(x)| \le C_2(T)p(x)(t-r)^{-(j-\theta)/2}\|f\|_{B([0,T];C_p^\theta(\mathbb{R}^n))}, \qquad j = 0,1,2. \tag{3.31}$$

To prove the previous estimate, it suffices to use the formula (3.23) and Lemma 3.10, recalling that $Q_0^{(k)}$ are strictly definite positive matrices, uniformly with respect to $t \in [0,T]$ and $k \in \mathbb{N}$. Then from (3.31), we deduce that

$$|D_t G_k(t,r)f(r,\cdot)(x)| \le (n^2 + nT^{1/2} + T)MC_2(T)p(x)\|f\|_{B([0,T];C_p^\theta(\mathbb{R}^n))}(t-r)^{-1+\theta/2}. \tag{3.32}$$

Therefore,

$$\begin{aligned} &|G_k(t_2,r)f(r,\cdot)(x) - G_k(t_1,r)f(r,\cdot)(x)| \\ &\le (n^2 + nT^{1/2} + T)MC_2(T)p(x)\|f\|_{B([0,T];C_p^\theta(\mathbb{R}^n))}(t_1-r)^{-1+\theta/2}|t_2-t_1| \end{aligned} \tag{3.33}$$

for any $0 \le r < t_1 \le t_2 \le T$. From (3.33), we easily deduce that $v(\cdot,x)$ is a Lipschitz continuous function in $[0,T]$. In fact, for any $0 \le t_1 \le t_2 \le T$, we have

$$\begin{aligned} |v_k(t_2,x) - v_k(t_1,x)| &\le \left| \int_{t_1}^{t_2} G_k(t_2,r)f(r,\cdot)(x)dr \right| \\ &\quad + \int_0^{t_1} \big|G_k(t_2,r)f(r,\cdot)(x) - G(t_1,r)f(r,\cdot)(x)\big|dr \\ &\le C_3(T)p(x)\|f\|_{B([0,T];C_p^\theta(\mathbb{R}^n))}|t_2-t_1|, \end{aligned} \tag{3.34}$$

$C_3(T)$ being a positive constant independent of (x,k). As k tends to infinity, we deduce that $v(\cdot,x)/p(x)$ is a Lipschitz function uniformly with respect to the variable x.

We are now in a position to prove that u/p is a Lipschitz continuous function in $[0,T] \times \mathbb{R}^n$. Suppose that t_1, $t_2 \in [0,T]$ and x_1, $x_2 \in \mathbb{R}^n$. Then

$$\begin{aligned} \left| \frac{u(t_2,x_2)}{p(x_2)} - \frac{u(t_1,x_1)}{p(x_1)} \right| &\le \left| \frac{u(t_2,x_2)}{p(x_2)} - \frac{u(t_1,x_2)}{p(x_2)} \right| + \left| \frac{u(t_1,x_2)}{p(x_2)} - \frac{u(t_1,x_1)}{p(x_1)} \right| \\ &\le \left[\frac{u(\cdot,x)}{p(x)} \right]_{\mathrm{Lip}\,([0,T])} |t_2-t_1| + \left\| D_x\left(\frac{u(t_1,\cdot)}{p} \right) \right\|_{L^\infty(\mathbb{R}^n)} |x_2-x_1|, \end{aligned} \tag{3.35}$$

and the assertion follows from (3.30), (3.34), and Theorems 3.2, 3.3, and 3.12. □

We are now in a position to prove the following existence theorem.

THEOREM 3.14. *Let $u_0 \in C_p^{2+\theta}(\mathbb{R}^n)$ with $0 < \theta < 1$ and let f be a measurable function belonging to $B([0,T];C_p^\theta(\mathbb{R}^n))$. The mild solution to* (3.26) *is twice continuously differentiable with respect to the space variables and it is differentiable with respect to t a.e. in $[0,T] \times \mathbb{R}^n$. Moreover, u is a solution to problem* (3.26) *in the sense of Definition* 1.1 *and*

$$\sup_{t\in[0,T]} \|u(t,\cdot)\|_{C_p^{2+\theta}(\mathbb{R}^n)} \le C\left(\|u_0\|_{C_p^{2+\theta}(\mathbb{R}^n)} + \sup_{t\in[0,T]} \|f(t,\cdot)\|_{C_p^\theta(\mathbb{R}^n)} \right) \tag{3.36}$$

for some positive constant C, independent of (u,u_0,f).

Proof. We start the proof considering the function v in Theorem 3.12. By Theorem 3.12 and Lemma 3.13, we already know that v is a continuous function in $[0,T] \times \mathbb{R}^n$ belonging to $B([0,T];C_p^{2+\theta}(\mathbb{R}^n))$. Therefore, $t \to v(t)$ belongs to $B([0,T];C^{2+\theta}(K)) \cap C([0,T];C(K))$ for any compact set $K \subset \mathbb{R}^n$ so that $v \in C([0,T];C^2(K))$, and hence v and its first- and second-order derivatives are continuous in $[0,T] \times \mathbb{R}^n$.

Thanks to Lemma 3.13, we deduce that $v(t,x)$ is differentiable a.e. in $[0,T]\times\mathbb{R}^n$ with respect to the variable t. Moreover,

$$D_t v(t,x) = \int_0^t D_t G(t,r) f(r,\cdot)(x)dr + f(t,x). \tag{3.37}$$

Let us prove that for any $x\in\mathbb{R}^n$, (3.37) holds in the sense of distributions. For this purpose, we observe that for any $\varphi\in C_0^\infty([0,T])$, we have (see also Lemmas A.1 and A.3 in the appendix)

$$\begin{aligned}
&\int_0^T \varphi'(t)dt\int_0^t G(t,r)f(r,\cdot)(x)dr\\
&\quad=\int_0^T dr\int_r^T \varphi'(t)G(t,r)f(r,\cdot)(x)dt\\
&\quad=-\int_0^T \varphi(r)f(r,x)dr-\int_0^T \varphi(t)dt\int_0^t D_tG(t,r)f(r,\cdot)(x)dr.
\end{aligned}$$

Here we have used the absolutely continuity of the function $t\to G(t,r)f(r,\cdot)(x)$ in $[r,T]$ and the measurability of the function $r\to f(r,x)$. To prove that $t\to G(t,r)f(r,\cdot)(x)$ is an absolutely continuous function in $[r,T]$ for any $(r,x)\in[0,T]\times\mathbb{R}^n$, it suffices to take (3.32) into account, observing that $t\to G_k(t,r)f(r,\cdot)(x)$ is an absolutely continuous function and $D_tG_k(t,r)f(r,\cdot)(x)$ tends to $D_tG(t,r)f(r,\cdot)(x)$ as $k\to+\infty$ a.e. in $[r,T]$.

Therefore, for any $x\in\mathbb{R}^n$, there exists a measurable set $F(x)$ with measure T such that

$$\begin{aligned}
D_t\int_0^t G(t,r)f(r,\cdot)(x)dr &= \int_0^t D_tG(t,r)f(r,\cdot)(x)dr + f(t,x)\\
&= \mathcal{A}(t)v(t,x)+f(t,x)
\end{aligned}\tag{3.38}$$

for any $t\in F(x)$.

Next we observe that the function

$$(t,x)\to\int_0^t G(t,r)f(r,\cdot)(x)dr$$

is measurable in $[0,T]\times\mathbb{R}^n$. Therefore, we can conclude that (3.37) holds a.e. in $[0,T]\times\mathbb{R}^n$. Then we consider the function $(t,x)\to G(t,0)u_0(x)$ and we observe that

$$D_xG(t,0)u_0(x)=G(t,0)D_xu_0(x),\qquad D_x^2G(t,0)u_0(x)=G(t,0)D_x^2u_0(x) \tag{3.39}$$

for any $(t,x)\in[0,T]\times\mathbb{R}^n$. Then it is easy to check that $G(\cdot,0)D_x^ju_0\in C_p([0,T]\times\mathbb{R}^n)$ $(j=0,1,2)$. From (3.23) and (3.28), we deduce that $u\in B([0,T];C_p^{2+\theta}(\mathbb{R}^n))$ and fulfills the estimate (3.36). □

To conclude this section, we show that the "mild" solution is the unique solution to problem (3.26).

THEOREM 3.15. *For any $u_0\in C_p^{2+\theta}(\mathbb{R}^n)$, $f\in B([0,T];C_p^\theta(\mathbb{R}^n))$, problem* (3.26) *admits a unique solution in the sense of Definition* 1.1.

Proof. Let us prove that problem (3.26) with $u_0\equiv 0$ and $f\equiv 0$ admits the trivial function as a unique solution. For this purpose, suppose that u is a solution in $[0,T]\times\mathbb{R}^n$ and fix a function $\varphi\in C^3(\mathbb{R}^n)$ with support in $B(0,1)$ and equal to

1 in $B(0,1/2)$. Then with any $x_0 \in \mathbb{R}^n$ we associate the function φ_{x_0} defined by $\varphi_{x_0}(x) = \varphi(x - x_0)$ and we define the function $v_{x_0} : [0,T] \times \mathbb{R}^n \to \mathbb{R}$ by $v_{x_0} = u\varphi_{x_0}$. As is easily seen, v_{x_0} is a Lipschitz continuous function in $[0,T] \times \mathbb{R}^n$. Moreover, it fulfills the Cauchy problem (3.26) in $[0,T] \times \mathbb{R}^n$ with $u_0 \equiv 0$ and f replaced by the function f_{x_0} given by

$$f_{x_0}(t,x) = -2\sum_{i,j=1}^n q_{i,j}(t)D_i u(t,x)D_j\varphi_{x_0}(x) - \sum_{i,j=1}^n q_{i,j}(t)u(t,x)D_{i,j}\varphi_{x_0}(x) - \sum_{j=1}^n b_j(t)u(t,x)D_j\varphi_{x_0}(x) \qquad \forall (t,x) \in [0,T] \times \mathbb{R}^n.$$

Next we observe that v_{x_0}, $f_{x_0} \in L^p([0,T] \times \mathbb{R}^n)$ for any $p \geq 1$. Consequently, we can apply the Fourier transform to the function $v_{x_0}(t,\cdot)$. As is easily checked, $\widehat{v}_{x_0}$ solves the following problem:

$$\begin{cases} D_t\widehat{v}_{x_0}(t,\xi) = \left(-\sum_{i,j=1}^n q_{i,j}(t)\xi_i\xi_j + \mathrm{i}\sum_{j=1}^n b_j(t)\xi_j + c(t)\right)\widehat{v}(t,\xi) + \widehat{f}_{x_0}, \\ \qquad\qquad (t,\xi) \in A \times \mathbb{R}^n, \\ \widehat{v}_{x_0}(0,\xi) = 0, \qquad \xi \in \mathbb{R}^n, \end{cases} \tag{3.40}$$

where A is a measurable set with measure T. Indeed, $\widehat{v}_{x_0}(\cdot,\xi)$ belongs to Lip $([0,T])$ for any $\xi \in \mathbb{R}^n$ with Lipschitz constant independent of ξ. In fact, $v_{x_0} \in$ Lip $([0,T] \times \mathbb{R}^n)$. Therefore, for any t_1, $t_2 \in [0,T]$,

$$|\widehat{v}_{x_0}(t_2,\xi) - \widehat{v}_{x_0}(t_1,\xi)| \leq \int_{B(x_0,1)} |v(t_2,x) - v(t_1,x)|dx \leq \omega_n[v]_{\mathrm{Lip}\,([0,T]\times\mathbb{R}^n)}|t_2 - t_1|,$$

where ω_n denotes, as usual, the Lebesgue measure of the unit ball in $\mathbb{R}^n$. Hence from the uniqueness of the solution to problem (3.40) in Lip $([0,T])$, we deduce that $\widehat{v}_{x_0}$ is given by the formula

$$\widehat{v}_{x_0}(t,\xi) = \int_0^t \exp\left(-\sum_{i,j=1}^n \int_s^t q_{i,j}(r)\xi_i\xi_j dr + \mathrm{i}\sum_{j=1}^n \xi_j\int_s^t b_j(r)dr + \int_s^t c(r)dr\right)\widehat{f}_{x_0}(s,\xi)ds$$

for any $(t,\xi) \in [0,T] \times \mathbb{R}^n$. Then taking the anti-Fourier transform of $\widehat{v}_{x_0}$, we get the following representation formula for v_{x_0}:

$$v_{x_0}(t,x) = \int_0^t G(t,r)f_{x_0}(r,\cdot)(x)dr \qquad \forall (t,x) \in [0,T] \times \mathbb{R}^n. \tag{3.41}$$

We now observe that f_{x_0} is a measurable function belonging to $B([0,l];C_p^\theta(\mathbb{R}^n))$ for any $0 < l \leq T$ and there exists a positive constant $C_1(p,T)$, independent of x_0, such that

$$\|f_{x_0}\|_{B([0,l];C_p^\theta(\mathbb{R}^n))} \leq C_1(p,T)\Big(\|u\|_{B([0,l];C_p^\theta(\mathbb{R}^n))} + \|u\|_{B([0,l];C_p^{1+\theta}(\mathbb{R}^n))}\Big). \tag{3.42}$$

Therefore, thanks to Theorem 3.12 we deduce that $v_{x_0} \in B([0,l];C_p^{2+\theta}(\mathbb{R}^n))$ for any $0 < l \leq T$ and

$$\|v_{x_0}\|_{B([0,l];C_p^{2+\theta}(\mathbb{R}^n))} \leq C_2(p,T)\Big(\|u\|_{B([0,l];C_p^\theta(\mathbb{R}^n))} + \|u\|_{B([0,l];C_p^{1+\theta}(\mathbb{R}^n))}\Big) \tag{3.43}$$

for some positive constant $C_2(p,T)$, independent of x_0. Recalling that $\varphi_{x_0} \equiv 1$ in $B(x_0,1/2)$, we deduce that $u \in B([0,T];C_p^\theta(B(x_0,1/2)))$ for any $x_0 \in \mathbb{R}^n$, and

$$\|u\|_{B([0,l];C_p^{2+\theta}(B(x_0,1/2)))} \le C_2(p,T)\Big(\|u\|_{B([0,l];C_p^\theta(\mathbb{R}^n))} + \|u\|_{B([0,l],C_p^{1+\theta}(\mathbb{R}^n))}\Big). \tag{3.44}$$

Since $C_2(p,T)$ is independent of $x_0 \in \mathbb{R}^n$, it is immediate to show that $u \in B([0,T];C_p^\theta(\mathbb{R}^n))$. Moreover, (3.44) can be extended to the whole of $\mathbb{R}^n$ by replacing the constant $C_2(p,T)$ with a new constant $C_3(p,T)$.

Then, taking advantage of [12, Proposition 1.1.3], we deduce that for any $\varepsilon > 0$, there exist two positive constants $K_1(\theta,\varepsilon)$ and $K_2(\theta,\varepsilon)$, independent of $l \in (0,T]$, such that

$$\|u\|_{B([0,l];C_p^\theta(\mathbb{R}^n))} \le \varepsilon\|u\|_{B([0,l];C_p^{2+\theta}(\mathbb{R}^n))} + K_1(\theta,\varepsilon)\|u\|_{B([0,l];C_p(\mathbb{R}^n))} \tag{3.45}$$

and

$$\|u\|_{B([0,l];C_p^{1+\theta}(\mathbb{R}^n))} \le \varepsilon\|u\|_{B([0,l];C_p^{2+\theta}(\mathbb{R}^n))} + K_2(\theta,\varepsilon)\|u\|_{B([0,l];C_p(\mathbb{R}^n))}. \tag{3.46}$$

Therefore, from (3.45) and (3.46), we deduce that there exists a positive constant $C_4(p,T)$, independent of x_0, such that

$$\|u\|_{B([0,l];C_p^{2+\theta}(\mathbb{R}^n))} \le C_4(p,T)\|u\|_{B([0,l];C_p(\mathbb{R}^n))}. \tag{3.47}$$

Then from Theorems 3.2 and 3.3, formula (3.41), estimates (3.42) and (3.47), and Lemma A.5 in the appendix, we deduce that

$$\left|\frac{u(s,x)}{p(x)}\right| \le C_5(p,T)\int_0^s \|u\|_{B([0,r];C_p(\mathbb{R}^n))}dr \qquad \forall (s,x) \in [0,T]\times B(x_0,1/2), \tag{3.48}$$

$C_5(p,T)$ being independent of x_0. Therefore, (3.48) can be extended to the whole of $[0,T]\times\mathbb{R}^n$ and

$$\|u\|_{B([0,t];C_p(\mathbb{R}^n))} \le C_5(p,T)\int_0^t \|u\|_{B([0,s];C_p(\mathbb{R}^n))}ds \qquad \forall t \in [0,T].$$

By means of Gronwall's inequality, we deduce that $u \equiv 0$ in $[0,T]\times\mathbb{R}^n$. □

4. The case of coefficients depending on (t,x). In this section, we are concerned with the following problem: *determine a function $u : [0,T]\times\mathbb{R}^n \to \mathbb{R}$ solution to the Cauchy problem*

$$\begin{cases} D_t u(t,x) = \displaystyle\sum_{i,j=1}^n q_{i,j}(t,x)D_{i,j}u(t,x) + \sum_{i=1}^n b_i(t,x)D_i u(t,x) \\ \qquad\qquad + c(t,x)u(t,x) + f(t,x), & (t,x) \in [0,T]\times\mathbb{R}^n, \\ u(0,x) = u_0(x), & x \in \mathbb{R}^n, \end{cases} \tag{4.1}$$

under the following assumptions on data:

(H1) $q_{i,j}$, b_j $(i,j = 1,\dots,n)$, and c belong to $L^\infty([0,T]\times\mathbb{R}^n) \cap B([0,T];C^\theta(\mathbb{R}^n))$ $(\theta \in (0,1))$;

(H2) there exists a positive constant C_0 such that $\sum_{i,j=1}^n q_{i,j}(t,x)\xi_i\xi_j \ge C_0$ for any $|\xi| = 1$ and any $(t,x) \in \mathcal{D}\times\mathbb{R}^n$, where $\mathcal{D}^c$ is a measurable set such that $\mathcal{D}^c \cap [0,T]$ is negligible;

(H3) $u_0 \in C_p^{2+\theta}(\mathbb{R}^n)$ and f is a measurable function belonging to $B([0,T];C_p^\theta(\mathbb{R}^n))$. For the sake of simplicity, we shall denote by $\mathcal{A}(t,x)$ the differential operator defined in (4.1), by $Q_0(t,x)$ the matrix having as elements the functions $q_{i,j}(t,x)$, by $B_0(t,x)$ the vector with components $(B_0)_j(t,x) = b_j(t,x)$, and by p either the polynomial or the exponential weight function.

4.1. A priori estimates and uniqueness of the solution. In this first part of section 4, we are interested in finding an a priori estimate for the solution to problem (4.1).

THEOREM 4.1. *Suppose that u is a solution to problem* (4.1) *in* $[0,T]\times\mathbb{R}^n$, *in the sense of Definition* 1.1, *belonging to* $B([0,T];C_p^{2+\theta}(\mathbb{R}^n))$. *Then u satisfies*

$$\sup_{t\in[0,T]} \|u(t,\cdot)\|_{C_p^{2+\theta}(\mathbb{R}^n)} \le C\left(\|u_0\|_{C_p^{2+\theta}(\mathbb{R}^n)} + \sup_{t\in[0,T]} \|f(t,\cdot)\|_{C_p^\theta(\mathbb{R}^n)}\right) \tag{4.2}$$

for some positive constant C, independent of (u,u_0,f).

Proof. Estimate (4.2) will be proved in two steps.

Step 1. Let us show that there exists a positive constant $\widetilde{C}$, independent of $T^*\in(0,T]$, such that

$$\|u\|_{B([0,T^*];C_p^{2+\theta}(\mathbb{R}^n))} \le \widetilde{C}\Big(\|u\|_{B([0,T^*];C_p(\mathbb{R}^n))} + \|u_0\|_{C_p^{2+\theta}(\mathbb{R}^n)} + \|f\|_{B([0,T^*];C_p^\theta(\mathbb{R}^n))}\Big) \tag{4.3}$$

for any $T^* \le T$. For this purpose, let $\psi \in C^3(\mathbb{R}^n)$ be such that $\psi \equiv 1$ in $B(0,1/2)$ with support in $B(0,1)$. With any $\delta > 0$ we associate the function $\psi_\delta \in C^3(\mathbb{R}^n)$ defined by $\psi_\delta(x) = \psi((x-x_0)/\delta)$ for any $x \in \mathbb{R}^n$. Then we consider the function $v_\delta = u\psi_\delta$ and observe that v_δ is a solution, in the sense of Definition 1.1, of the following Cauchy problem:

$$\begin{cases} D_t v_\delta(t,x) = \mathcal{A}(t,x)v_\delta(t,x) + f(t,x) - g_\delta(t,x), & (t,x)\in[0,T]\times\mathbb{R}^n, \\ v_\delta(0,x) = u_0(x)\psi_\delta(x), & x\in\mathbb{R}^n, \end{cases} \tag{4.4}$$

where

$$\begin{aligned} g_\delta(t,x) = 2\sum_{i,j=1}^n q_{i,j}(t,x)D_i u(t,x)D_j\psi_\delta(x) + \sum_{i,j=1}^n q_{i,j}(t,x)u(t,x)D_{i,j}^2\psi_\delta(x) \\ +\sum_{i=1}^n b_i(t,x)u(t,x)D_i\psi_\delta(x) \qquad \forall (t,x)\in[0,T^*]\times\mathbb{R}^n. \end{aligned} \tag{4.5}$$

Then we fix $x_0 \in \mathbb{R}^n$ and rewrite problem (4.4) in the following form:

$$\begin{cases} D_t v_\delta(t,x) = \mathcal{A}(t,x_0)v_\delta(t,x) + \big(\mathcal{A}(t,x) - \mathcal{A}(t,x_0)\big)v_\delta(t,x) \\ \qquad\qquad + f(t,x) - g_\delta(t,x), & (t,x)\in[0,T]\times\mathbb{R}^n, \\ v_\delta(0,x) = u_0(x)\psi_\delta(x), & x\in\mathbb{R}^n. \end{cases} \tag{4.6}$$

An easy computation shows that $u_0\psi_\delta \in C_p^{2+\theta}(\mathbb{R}^n)$ and

$$\|u_0\psi_\delta\|_{C_p^{2+\theta}(\mathbb{R}^n)} \le C(\delta)\|u_0\|_{C_p^{2+\theta}(\mathbb{R}^n)}, \tag{4.7}$$

$C(\delta)$ being a positive constant, independent of x_0, tending to $+\infty$ as $\delta \to 0$. We now prove that the function $k = \mathcal{A}(\cdot,\cdot)v_\delta - \mathcal{A}(\cdot,x_0)v_\delta$ belongs to $B([0,T];C_p^\theta(\mathbb{R}^n))$. For this purpose, we observe that for any x_1, $x_2 \in B(x_0,\delta)$, $y \in \partial B(x_0,\delta)$, any i, $j \in \mathbb{N}$, $1 \le i,j \le n$, and any $t \in [0,T]$,

$$\begin{aligned}
&\left|\big[q_{i,j}(t,x_2)-q_{i,j}(t,x_0)\big]\frac{D_{i,j}v_\delta(t,x_2)}{p(x_2)} - \big[q_{i,j}(t,x_1)-q_{i,j}(t,x_0)\big]\frac{D_{i,j}v_\delta(t,x_1)}{p(x_1)}\right| \\
&\quad \le |q_{i,j}(t,x_2)-q_{i,j}(t,x_1)|\left|\frac{D_{i,j}v_\delta(t,x_2)}{p(x_2)} - \frac{D_{i,j}v_\delta(t,y)}{p(y)}\right| \\
&\qquad + |q_{i,j}(t,x_1)-q_{i,j}(t,x_0)|\left|\frac{D_{i,j}v_\delta(t,x_2)}{p(x_2)} - \frac{D_{i,j}v_\delta(t,x_1)}{p(x_1)}\right| \\
&\quad \le 3\delta^\theta [q_{i,j}(t,\cdot)]_{C^\theta(\mathbb{R}^n)}\left[\frac{D_{i,j}v_\delta(t,\cdot)}{p}\right]_{C^\theta(\mathbb{R}^n)}|x_2-x_1|^\theta. \qquad (4.8)
\end{aligned}$$

Suppose now that $x_2 \in B(x_0,\delta)$, $x_1 \in \mathbb{R}^n\backslash B(x_0,\delta)$. Then

$$\begin{aligned}
&\left|\big[q_{i,j}(t,x_2)-q_{i,j}(t,x_0)\big]\frac{D_{i,j}v_\delta(t,x_2)}{p(x_2)} - \big[q_{i,j}(t,x_1)-q_{i,j}(t,x_0)\big]\frac{D_{i,j}v_\delta(t,x_1)}{p(x_1)}\right| \\
&\quad = \left|\big[q_{i,j}(t,x_2)-q_{i,j}(t,x_0)\big]\left[\frac{D_{i,j}v_\delta(t,x_2)}{p(x_2)} - \frac{D_{i,j}v_\delta(t,x_1)}{p(x_1)}\right]\right| \\
&\quad \le \delta^\theta [q_{i,j}(t,\cdot)]_{C^\theta(\mathbb{R}^n)}\left[\frac{D_{i,j}v_\delta(t,\cdot)}{p}\right]_{C^\theta(\mathbb{R}^n)}|x_2-x_1|^\theta. \qquad (4.9)
\end{aligned}$$

Moreover,

$$\big\|\big[q_{i,j}(t,\cdot)-q_{i,j}(t,x_0)\big]D_{i,j}v_\delta(t,\cdot)\big\|_{C_p(\mathbb{R}^n)} \le \delta^\theta [q_{i,j}(t,\cdot)]_{C^\theta(\mathbb{R}^n)}\|D_{i,j}v_\delta(t,\cdot)\|_{C_p(\mathbb{R}^n)}. \qquad (4.10)$$

From (4.8)–(4.10), we deduce that

$$\big\|\big[q_{i,j}(t,\cdot)-q_{i,j}(t,x_0)\big]v_\delta(t,\cdot)\big\|_{C_p^\theta(\mathbb{R}^n)} \le 3\delta^\theta [q_{i,j}(t,\cdot)]_{C^\theta(\mathbb{R}^n)}\|D_{i,j}v_\delta(t,\cdot)\|_{C_p^\theta(\mathbb{R}^n)}. \qquad (4.11)$$

In the same way, it can be proved that

$$\big\|\big[b_j(t,\cdot)-b_j(t,x_0)\big]D_jv_\delta(t,\cdot)\big\|_{C_p^\theta(\mathbb{R}^n)} \le 3\delta^\theta [b_j(t,\cdot)]_{C^\theta(\mathbb{R}^n)}\|D_jv_\delta(t,\cdot)\|_{C_p^\theta(\mathbb{R}^n)} \qquad (4.12)$$

for any $j = 1,\dots,n$ and

$$\big\|\big[c(t,\cdot)-c(t,x_0)\big]v_\delta(t,\cdot)\big\|_{C_p^\theta(\mathbb{R}^n)} \le 3\delta^\theta [c(t,\cdot)]_{C^\theta(\mathbb{R}^n)}\|v_\delta(t,\cdot)\|_{C_p^\theta(\mathbb{R}^n)}. \qquad (4.13)$$

Then, recalling that for any $\varphi \in C^1(\mathbb{R}^n)$, $\|\varphi\|_{C^\theta(\mathbb{R}^n)} \le 3\|\varphi\|_{C^1(\mathbb{R}^n)}$, and taking Lemma 2.5 into account, we deduce that $k \in B([0,T];C_p^\theta(\mathbb{R}^n))$ and there exists a positive constant $C(T)$, independent of x_0, δ and $T^* \in (0,T]$, such that

$$\|k\|_{B([0,T^*];C_p^\theta(\mathbb{R}^n))} \le C(T)\delta^\theta \|v_\delta\|_{B([0,T^*];C_p^{2+\theta}(\mathbb{R}^n))}. \qquad (4.14)$$

As far as g_δ is concerned, we observe that it belongs to $B([0,T];C_p^\theta(\mathbb{R}^n))$ for any $\delta > 0$, and if $\delta \le 1$, $t \in [0,T]$,

$$\|g_\delta(t,\cdot)\|_{C_p^\theta(\mathbb{R}^n)} \le \delta^{-2-\theta}\Bigg[2\sum_{i,j=1}^n \|q_{i,j}(t,\cdot)\|_{C^\theta(\mathbb{R}^n)}\|D\psi\|_{C^\theta(\mathbb{R}^n)}\|u(t,\cdot)\|_{C_p^{1+\theta}(\mathbb{R}^n)} + \sum_{i,j=1}^n \|q_{i,j}(t,\cdot)\|_{C^\theta(\mathbb{R}^n)}\|u(t,\cdot)\|_{C_p^\theta(\mathbb{R}^n)}\|D^2\psi\|_{C^\theta(\mathbb{R}^n)} + \sum_{j=1}^n \|b_j(t,\cdot)\|_{C^\theta(\mathbb{R}^n)}\|u(t,\cdot)\|_{C_p^\theta(\mathbb{R}^n)}\|D\psi\|_{C^\theta(\mathbb{R}^n)}\Bigg]. \tag{4.15}$$

Thanks to Theorems 3.14 and 3.15 and estimates (4.14) and (4.15), we can find a positive constant $D_0(T)$, independent of x_0 and δ, such that

$$\|v_\delta\|_{B([0,T^*];C_p^{2+\theta}(\mathbb{R}^n))} \le D_0(T)\Big(C(\delta)\|u_0\|_{C_p^{2+\theta}(\mathbb{R}^n)} + C(T)\delta^\theta\|v_\delta\|_{B([0,T^*];C_p^{2+\theta}(\mathbb{R}^n))} + \|g_\delta\|_{B([0,T^*];C_p^\theta(\mathbb{R}^n))} + \|f\|_{B([0,T^*];C_p^\theta(\mathbb{R}^n))}\Big). \tag{4.16}$$

Therefore, for any $\delta \le \delta_0 = \min\,(1,(D_0(T)C(T))^{-1})$, recalling that $v_\delta \equiv u$ in $B(x_0,\delta/2)$, we deduce that

$$\|u\|_{B([0,T^*];C_p^{2+\theta}(B(x_0,\delta/2)))} \le D_1(T,\delta)\Big(\|u_0\|_{C_p^{2+\theta}(\mathbb{R}^n)} + \|f\|_{B([0,T^*];C_p^\theta(\mathbb{R}^n))} + \|g_\delta\|_{B([0,T^*];C_p^\theta(\mathbb{R}^n))}\Big), \tag{4.17}$$

where $D_1(T,\delta) = \frac{D_0(T)\max\,(C(\delta),1)}{1-D_0(T)C(T)\delta^\theta}$. A direct inspection of (4.15) shows that (4.17) is independent of the point x_0. Therefore,

$$\|u\|_{B([0,T^*];C_p^2(\mathbb{R}^n))} \le D_1(T,\delta)\Big(\|u_0\|_{C_p^{2+\theta}(\mathbb{R}^n)} + \|f\|_{B([0,T^*];C_p^\theta(\mathbb{R}^n))} + \|g_\delta\|_{B([0,T^*];C_p^\theta(\mathbb{R}^n))}\Big). \tag{4.18}$$

Moreover, it can be easily proved that for any $t \in [0,T]$ and any $\delta \le \delta_0$

$$\left[D_{i,j}\frac{u(t,\cdot)}{p}\right]_{C^\theta(\mathbb{R}^n)} \le 2^{\theta+1}\delta^{-\theta}D_1(T,\delta) \times \Big(\|u_0\|_{C_p^{2+\theta}(\mathbb{R}^n)} + \|f\|_{B([0,T^*];C_p^\theta(\mathbb{R}^n))} + \|g_\delta\|_{B([0,T^*];C_p^\theta(\mathbb{R}^n))}\Big). \tag{4.19}$$

Consequently, for any $T^* \in (0,T]$,

$$\|u\|_{B([0,T^*];C_p^{2+\theta}(\mathbb{R}^n))} \le D_2(T,\delta)\Big(\|u_0\|_{C_p^{2+\theta}(\mathbb{R}^n)} + \|f\|_{B([0,T^*];C_p^\theta(\mathbb{R}^n))} + \|g_\delta\|_{B([0,T^*];C_p^\theta(\mathbb{R}^n))}\Big), \tag{4.20}$$

where $D_2(T,\delta) = \big(2^{\theta+1}\delta^{-\theta}n^2+1\big)D_1(T,\delta)$. Then, taking (3.45), (3.46), and Lemma 2.5 into account, from (4.15) and (4.20), we deduce that there exist two positive constants $D_3(T)$ and $K(\theta,\varepsilon)$ such that for any $\delta \le 1$,

$$\|g_\delta\|_{B([0,T^*];C_p^\theta(\mathbb{R}^n))} \le \delta^{-2-\theta}D_3(T)\Big(\varepsilon\|u\|_{B([0,T^*];C_p^{2+\theta}(\mathbb{R}^n))} + K(\theta,\varepsilon)\|u\|_{B([0,T^*];C_p(\mathbb{R}^n))}\Big). \tag{4.21}$$

From (4.20) and (4.21), we easily deduce that for ε sufficiently small and $\delta \le \delta_0$,

$$\begin{aligned}&\|u\|_{B([0,T^*];C_p^{2+\theta}(\mathbb{R}^n))}\\ &\le D_4(T,\delta,\varepsilon)\Big(\|u_0\|_{C_p^{2+\theta}(\mathbb{R}^n)}+\|u\|_{B([0,T^*];C_p(\mathbb{R}^n))}+\|f\|_{B([0,T^*];C_p^{\theta}(\mathbb{R}^n))}\Big),\end{aligned}\tag{4.22}$$

$D_4(T,\delta,\varepsilon)$ being a positive constant, and (4.3) follows.

Step 2. Let us fix $x_0 \in \mathbb{R}^n$ and rewrite problem (4.1) in the following form:

$$\begin{cases} D_t u(t,x) = \mathcal{A}(t,x_0)u(t,x) + \big(\mathcal{A}(t,x)-\mathcal{A}(t,x_0)\big)u(t,x) + f(t,x),\\ \qquad\qquad (t,x)\in[0,T]\times\mathbb{R}^n,\\ u(0,x)=u_0(x), \qquad x\in\mathbb{R}^n. \end{cases}\tag{4.23}$$

Then an easy computation shows that $g(t,x)=(\mathcal{A}(t,x)-\mathcal{A}(t,x_0))u(t,x)$ belongs to $B([0,T];C_p^{\theta}(\mathbb{R}^n))$ with norm independent of x_0. Therefore, thanks to Theorems 3.14 and 3.15, we deduce that u admits the following representation:

$$u(t,x)=G(t,0)u_0(x)+\int_0^t G(t,s)\big[\big(\mathcal{A}(s,\cdot)-\mathcal{A}(s,x_0)\big)u(s,\cdot)+f(s,\cdot)\big](x)ds.\tag{4.24}$$

Then for any $s\in[0,t]$,

$$\begin{aligned}&\left|\frac{\mathcal{A}(s,x)u(s,x)}{p(x)}-\frac{\mathcal{A}(s,x_0)u(t,x)}{p(x)}\right|\\ &\quad\le\Bigg[\sum_{i,j=1}^n \|q_{i,j}\|_{B([0,s];C^{\theta}(\mathbb{R}^n))}\|D_x^2u\|_{B([0,s];C_p^{\theta}(\mathbb{R}^n))}\\ &\qquad+\sum_{j=1}^n\|b_j\|_{B([0,s];C^{\theta}(\mathbb{R}^n))}\|D_xu\|_{B([0,s];C_p^{\theta}(\mathbb{R}^n))}\\ &\qquad+\|c\|_{B([0,s];C^{\theta}(\mathbb{R}^n))}\|u\|_{B([0,s];C_p(\mathbb{R}^n))}\Bigg]|x-x_0|^{\theta}\\ &\quad\le C_1(T)\Big(\|u\|_{B([0,s];C_p(\mathbb{R}^n))}+\|u_0\|_{C_p^{2+\theta}(\mathbb{R}^n)}+\|f\|_{B([0,s];C_p^{2+\theta}(\mathbb{R}^n))}\Big)|x-x_0|^{\theta},\end{aligned}\tag{4.25}$$

where $C_1(T)$ is a positive constant independent of x_0. Therefore, from (3.5), (3.17), and (4.25), we deduce that there exists a positive constant $C_2(T)$, independent of x_0, such that for any $x\in B(x_0,1)$,

$$\begin{aligned}&\left|\frac{G(t,s)\big(\mathcal{A}(s,x)-\mathcal{A}(s,x_0)\big)u(s,\cdot)(x)}{p(x)}\right|\\ &\qquad\le C_2(T)\Big(\|u\|_{B([0,s];C_p(\mathbb{R}^n))}+\|u_0\|_{C_p^{2+\theta}(\mathbb{R}^n)}+\|f\|_{B([0,s];C_p^{2+\theta}(\mathbb{R}^n))}\Big).\end{aligned}\tag{4.26}$$

Then taking advantage of (4.24), (4.26), and Lemma A.5 in the appendix, we deduce that

$$\left|\frac{u(t,x)}{p(x)}\right|\le C_3(T)\left[\|u_0\|_{C_p^{2+\theta}(\mathbb{R}^n)}+\int_0^t\|u\|_{B([0,s];C_p(\mathbb{R}^n))}ds+\|f\|_{B([0,t];C_p^{\theta}(\mathbb{R}^n))}\right]\tag{4.27}$$

for some positive constant $C_3(T)$, independent of x_0, and for any $(t,x) \in [0,T] \times B(x_0,1)$. Then we extend the previous estimate to the whole of $\mathbb{R}^n$ by observing that (4.27) is independent of x_0. By means of Gronwall's inequality, we deduce that there exists a positive constant $C_4(T)$ such that

$$\|u\|_{B([0,t];C_p(\mathbb{R}^n))} \le C_4(T)\Big(\|u_0\|_{C_p^{2+\theta}(\mathbb{R}^n)} + \|f\|_{B([0,t];C_p^\theta(\mathbb{R}^n))}\Big) \qquad \forall t \in [0,T]. \tag{4.28}$$

From (4.3) and (4.28), we deduce (4.2). □

4.2. Existence of the solution. Now we are in a position to prove that problem (4.1) admits a solution. For this purpose, we use the classical method of continuity.

THEOREM 4.2. *For any $u_0 \in C_p^{2+\theta}(\mathbb{R}^n)$ and any measurable function f belonging to $B([0,T];C_p^\theta(\mathbb{R}^n))$, problem (4.1) admits a unique solution in the sense of Definition 1.1. Moreover, there exists a positive constant C such that*

$$\|u\|_{B([0,T];C_p^\theta(\mathbb{R}^n))} \le C\Big(\|u_0\|_{C_p^{2+\theta}(\mathbb{R}^n)} + \|f\|_{B([0,T];C_p^\theta(\mathbb{R}^n))}\Big). \tag{4.29}$$

Proof. With any $\lambda \in [0,1]$ we associate the differential operator $\mathcal{A}_\lambda = \lambda\mathcal{A} + (1-\lambda)\Delta$. As is easily seen, the coefficients of $\mathcal{A}_\lambda$ are bounded in $B([0,T];C^\theta(\mathbb{R}^n))$ by positive constants independent of λ. Moreover (cf. assumption H2),

$$\langle \lambda Q(t,x)\xi + (1-\lambda)\xi, \xi\rangle \ge \min(C_0,1)|\xi|^2 \qquad \forall (t,x) \in \mathcal{D} \times \mathbb{R}^n.$$

Next we denote by $\mathcal{F}$ the set of all the $\lambda \in [0,1]$ such that for any $u_0 \in C_p^{2+\theta}(\mathbb{R}^n)$ and any function f belonging to $B([0,T];C_p^\theta(\mathbb{R}^n))$, (P_λ) admits a solution where

$$(\mathrm{P}_\lambda) \quad \begin{cases} D_t u(t,x) = \mathcal{A}_\lambda u(t,x) + f(t,x), & (t,x) \in [0,T] \times \mathbb{R}^n, \\ u(0,x) = u_0(x), & x \in \mathbb{R}^n. \end{cases} \tag{4.30}$$

Taking advantage of Theorem 4.1, we deduce that there exists a positive constant C_1, independent of λ, such that if u_λ is a measurable solution to (P_λ), then

$$\|u_\lambda\|_{B([0,T];C_p^{2+\theta}(\mathbb{R}^n))} \le C_1\Big(\|u_0\|_{C_p^{2+\theta}(\mathbb{R}^n)} + \|f\|_{B([0,T];C_p^\theta(\mathbb{R}^n))}\Big). \tag{4.31}$$

We observe that by section 3, $\lambda = 0$ belongs to $\mathcal{F}$. Now we prove that $\mathcal{F}$ is a closed subset of $[0,1]$. For this purpose suppose that $\{\lambda_n\}_{n\in\mathbb{N}} \subset \mathcal{F}$ tends to λ as $n \to +\infty$. Let u_{λ_n} be the solution to problem (P_{λ_n}). Then $u_{\lambda_n} - u_{\lambda_m}$ turns out to be a solution to the Cauchy problem

$$\begin{cases} D_t\big(u_{\lambda_n} - u_{\lambda_m}\big)(t,x) = \mathcal{A}_{\lambda_n}(t,x)\Big(u_{\lambda_n}(t,x) - u_{\lambda_m}(t,x)\Big) \\ \qquad\qquad + \big(\mathcal{A}_{\lambda_n} - \mathcal{A}_{\lambda_m}\big)(t,x)u_{\lambda_m}(t,x), \qquad (t,x) \in [0,T] \times \mathbb{R}^n, \\ \big(u_{\lambda_n} - u_{\lambda_m}\big)(0,x) = 0, \qquad x \in \mathbb{R}^n. \end{cases} \tag{4.32}$$

Taking advantage of Lemma 2.5 and (4.31), it can be easily proved that $(\mathcal{A}_{\lambda_n} - \mathcal{A}_{\lambda_m})u_{\lambda_m}$ is measurable, belongs to $B([0,T];C_p^\theta(\mathbb{R}^n))$, and

$$\begin{aligned} \|\big(\mathcal{A}_{\lambda_n} - \mathcal{A}_{\lambda_m}\big)u_{\lambda_m}\|_{C_p^\theta(\mathbb{R}^n)} &\le C(p)|\lambda_n - \lambda_m|\Big(\|u_0\|_{C_p^{2+\theta}(\mathbb{R}^n)} + \|f\|_{B([0,T];C_p^\theta(\mathbb{R}^n))}\Big) \\ &\times \left(\sum_{i,j=1}^n \|q_{i,j}\|_{B([0,T];C^\theta(\mathbb{R}^n))} + \sum_{j=1}^n \|b_j\|_{B([0,T];C^\theta(\mathbb{R}^n))} + \|c\|_{B([0,T];C^\theta(\mathbb{R}^n))}\right). \end{aligned} \tag{4.33}$$

From (4.33), we deduce that $\{u_{\lambda_n}\}_{n\in\mathbb{N}}$ is a Cauchy sequence in $B([0,T];C_p^{2+\theta}(\mathbb{R}^n))$. Therefore, there exists $u \in B([0,T];C_p^{2+\theta}(\mathbb{R}^n))$ such that $u_{\lambda_n} \to u$ as $n \to +\infty$. We remark here that u_{λ_n}, $D_x u_{\lambda_n}$, $D_x^2 u_{\lambda_n} \in C([0,T]\times\mathbb{R}^n)$. Hence u and its first- and second-order space derivatives are continuous functions in $[0,T]\times\mathbb{R}^n$. Moreover, $D_t(u_{\lambda_n} - u_{\lambda_m})$ is a Cauchy sequence in $L^\infty(G)$, where G is a measurable subset of $[0,T]\times\mathbb{R}^n$ with negligible complement in $[0,T]\times\mathbb{R}^n$ and such that for any $x\in\mathbb{R}^n$, the set $G(x) = \{t\in[0,T] : (t,x)\in G\}$ is measurable with measure T and $D_t u_{\lambda_n} = \mathcal{A}(\cdot,\cdot)u_{\lambda_n} + f$ in G. Therefore, there exists a measurable function h such that $D_t u_{\lambda_n} \to h$ in $L^\infty(G)$. By assumption, $u_{\lambda_n}(\cdot,x)$, $x\in\mathbb{R}^n$, is Lipschitz continuous in $[0,T]$. Therefore, from (4.31) and (4.33), we deduce that

$$\begin{aligned}\Big|\frac{u_{\lambda_n}(t_2,x)}{p(x)} - \frac{u_{\lambda_n}(t_1,x)}{p(x)}\Big| &\le \int_{t_1}^{t_2}\frac{|D_t u_{\lambda_n}(s,x)|}{p(x)}ds\\ &\le C_2\Big(\|u_0\|_{C_p^{2+\theta}(\mathbb{R}^n)} + \|f\|_{B([0,T];C_p^\theta(\mathbb{R}^n))}\Big)|t_2-t_1| \qquad (4.34)\end{aligned}$$

for any t_1, $t_2\in[0,T]$, any $x\in\mathbb{R}^n$, and some positive constant C_2. As $n\to+\infty$ we deduce that $u(\cdot,x)$ is a Lipschitz continuous function for any $x\in\mathbb{R}^n$. Then, reasoning as in (3.35), we can prove that $u/p \in \mathrm{Lip}\,([0,T]\times\mathbb{R}^n)$. Now it is an easy task to show that for any $x\in\mathbb{R}^n$, there exists a measurable set $H(x)$ with measure T such that $h(t,x) = D_t u(t,x)$ and $D_t u(t,x) = \mathcal{A}(t,x)u(t,x) + f(t,x)$ for any $t\in H(x)$. Next we observe that $D_t u$, h, and $\mathcal{A}(\cdot,\cdot)u + f$ are a.e. defined and measurable in $[0,T]\times\mathbb{R}^n$. Therefore, thanks to the Fubini–Tonelli theorem, it can be proved that $D_t u = h$ and $D_t u = \mathcal{A}(\cdot,\cdot)u + f$ a.e. in $[0,T]\times\mathbb{R}^n$ so that condition (iii) in Definition 1.1 holds. Consequently, $\mathcal{F}$ is closed in $[0,1]$.

Now we prove that $\mathcal{F}$ is open in $[0,1]$. Thanks to (4.31), for any $\lambda\in\mathcal{F}$, we can define the operator $\mathcal{M}(\lambda)$ that with any pair (u_0,f) of initial data associates the solution to problem (P_λ). Suppose that $\lambda_0\in\mathcal{F}$ and consider λ next to λ_0. Then we write (P_λ) in the following form:

$$\begin{cases} D_t u(t,x) = \mathcal{A}_{\lambda_0}(t,x)u(t,x) + \big(\mathcal{A}_\lambda - \mathcal{A}_{\lambda_0}\big)(t,x)u(t,x) + f(t,x),\\ \qquad\qquad (t,x)\in[0,T]\times\mathbb{R}^n,\\ u(0,x) = u_0(x), \qquad x\in\mathbb{R}^n. \end{cases} \qquad (4.35)$$

As is easily seen, for any $u\in B([0,T];C_p^{2+\theta}(\mathbb{R}^n))$, $(\mathcal{A}_\lambda - \mathcal{A}_{\lambda_0})u$ is a measurable function belonging to $B([0,T];C_p^\theta(\mathbb{R}^n))$ and satisfying estimate (4.33) with (λ_n,λ_m) replaced by (λ,λ_0) and u_{λ_n} replaced by u. Therefore, we deduce that if u is a measurable solution to (P_λ), then u solves the following fixed-point problem

$$u = F(u) := \mathcal{M}(\lambda_0)(u_0,0) + \mathcal{M}(\lambda_0)(0,\big(\mathcal{A}_\lambda - \mathcal{A}_{\lambda_0}\big)u) + M(\lambda_0)(0,f). \qquad (4.36)$$

F is a contraction map in $\mathcal{X} = \{f\in B([0,T];C_p^{2+\theta}(\mathbb{R}^n)) : f \text{ is measurable in } [0,T]\times\mathbb{R}^n\}$ that is a Banach space when endowed with the norm of $B([0,T];C_p^{2+\theta}(\mathbb{R}^n))$. In fact, F maps $\mathcal{X}$ into itself. Moreover, for any u_1, $u_2\in\mathcal{X}$,

$$\begin{aligned}\|F(u_2) - F(u_1)\|_{B([0,T];C_p^{2+\theta}(\mathbb{R}^n))} &= \|\mathcal{M}(\lambda_0)(0,\mathcal{A}_\lambda - \mathcal{A}_{\lambda_0})(u_2-u_1)\|_{B([0,T];C_p^{2+\theta}(\mathbb{R}^n))}\\ &\le C_3(p,Q_0,B_0,c)|\lambda-\lambda_0|\|u_1-u_2\|_{B([0,T];C_p^{2+\theta}(\mathbb{R}^n))}\end{aligned}$$

for some positive constant $C_3(p,Q_0,B_0,c)$. Therefore, for $|\lambda-\lambda_0|$ sufficiently small, F is a contraction map in $B([0,T];C_p^{2+\theta}(\mathbb{R}^n))$, and consequently (4.36) admits a unique

solution $u \in B([0,T];C_p^{2+\theta}(\mathbb{R}^n))$. Then it is easy to show that such u is a solution to (P_λ). Therefore, $\lambda \in \mathcal{F}$. Being an open and closed set, $\mathcal{F}$ coincides with $[0,1]$ and the assertion follows. □

Appendix.

LEMMA A.1. *For any $f : [0,T] \times \mathbb{R}^n \to \mathbb{R}$ such that $f/p \in L^\infty([0,T] \times \mathbb{R}^n)$ and $f(r,\cdot)$ is measurable for any $r \in [0,T]$, we define the functions $L(f)$, $M(f) : [0,T]^2 \times \mathbb{R}^n \to \mathbb{R}$ as follows:*

$$L(f)(t,r,x) = \begin{cases} G(t,r)f(r,\cdot)(x), & (t,r,x) \in E(T) \times \mathbb{R}^n, \\ 0 & \text{elsewhere in } [0,T]^2 \times \mathbb{R}^n, \end{cases}$$

and

$$M(f)(t,r,x) = \begin{cases} D_t G(t,r)f(r,\cdot)(x), & (t,r,x) \in \big(\widetilde{E}(T) \cap (E_0 \times [0,T])\big) \times \mathbb{R}^n, \\ 0 & \text{elsewhere in } [0,T]^2 \times \mathbb{R}^n, \end{cases}$$

where $E(T) = \{(t,r) \in [0,T]^2 : r \le t\}$, $\widetilde{E}(T) = \{(t,r) \in [0,T]^2 : r < t,\ t \in E_0\}$, and E_0 is defined in Lemma 3.7. Then the following properties hold true:

(i) *$L(f)$ and $M(f)$ are measurable in $[0,T]^2 \times \mathbb{R}^n$;*
(ii) *for any $x \in \mathbb{R}^n$, the functions $L(f)(\cdot,\cdot,x)$ and $M(f)(\cdot,\cdot,x)$ are measurable in $[0,T]^2$;*
(iii) *for any $(t,x) \in [0,T] \times \mathbb{R}^n$, the functions $L(f)(t,\cdot,x)$ and $M(f)(t,\cdot,x)$ are measurable in $[0,T]$;*
(iv) *for any $(r,x) \in [0,T] \times \mathbb{R}^n$, the functions $L(f)(\cdot,r,x)$ and $M(f)(\cdot,r,x)$ are measurable in $[0,T]$;*
(v) *the function $N(f) : E_0 \times \mathbb{R}^n \to \mathbb{R}$ defined by*

$$N(f)(t,x) = \int_0^t D_t G(t,r)f(r,\cdot)(x)dr$$

for any $(t,x) \in E_0 \times \mathbb{R}^n$, is measurable.

Proof. We sketch the proof. Let us consider $L(f)$ and suppose that f is a continuous function in $[0,T] \times \mathbb{R}^n$ with $f/p \in L^\infty([0,T] \times \mathbb{R}^n)$. Then $L(f)$ is a continuous function in $E(T)$. Hence it is measurable in $[0,T]^2 \times \mathbb{R}^n$. Next we suppose that f is a measurable function with support contained in $[0,T] \times B(0,h)$ for some $h \in [0,T]$. From Lusin's theorem, we deduce that there exists a sequence $\{f_k\}_{k\in\mathbb{N}}$ of continuous functions converging to f a.e. in $[0,T] \times \mathbb{R}^n$ and such that $\|f_k/p\|_\infty \le \|f/p\|_\infty$ for any $k \in \mathbb{N}$. Then an easy application of the dominated convergence theorem shows that the sequence $\{L(f_k)\}_{k\in\mathbb{N}}$ converges to $L(f)$ as $k \to +\infty$ in $E(T) \cap ([0,T] \times \mathcal{C}) \times \mathbb{R}^n$. Here $\mathcal{C}$ denotes the set of all $r \in [0,T]$ such that $f_k(r,\cdot) \to f(r,\cdot)$ a.e. in $\mathbb{R}^n$ as $k \to +\infty$. Obviously, $\mathcal{C}$ is a measurable set in $[0,T]$ with measure T. Hence $L(f)$ is still a measurable function. Next, with any measurable function f such that $f/p \in L^\infty([0,T] \times \mathbb{R}^n)$, we associate the sequence $\{f_k\}_{k\in\mathbb{N}}$ defined by $f_k \equiv f$ in $[0,T] \times B(0,k)$ and $f_k \equiv 0$ elsewhere in $[0,T] \times \mathbb{R}^n$. $L(f_k)$ is measurable and converges to $L(f)$ everywhere in $[0,T]^2 \times \mathbb{R}^n$. Hence $L(f)$ is measurable in $[0,T]^2 \times \mathbb{R}^n$.

The same technique can be used to prove that for any $t \in [0,T]$ and any $x \in \mathbb{R}^n$, the functions $L(f)(t,\cdot,x)$, $L(f)(\cdot,\cdot,x)$ are measurable in $[0,T]$ and $[0,T]^2$, respectively. Moreover, to prove that $t \to Lf(t,r,x)$ is measurable in $[0,T]$ for any $(r,x) \in [0,T] \times \mathbb{R}^n$, it suffices to observe that the function $t \to G(t,r)f(r,\cdot)(x)$ is continuous in $(r,T]$.

We now consider the function $M(f)$. We observe that for any continuous function f such that f/p is bounded in $[0,T]\times\mathbb{R}^n$, then $M(f)$ is measurable in $[0,T]^2\times\mathbb{R}^n$. In fact, let us consider the approximating functions $Q_0^{(k)}$, $B_0^{(k)}$, and $c^{(k)}$ defined in Lemma 3.10 and the semigroup G_k defined as G with Q_0, B_0, and c replaced by $Q_0^{(k)}$, $B_0^{(k)}$, and $c^{(k)}$, respectively. Then it is easy to check that the function $(t,r,x)\to D_tG_k(t,r)f(r,\cdot)(x)$ is continuous in $\widetilde{E}(T)\times\mathbb{R}^n$. Moreover, $D_tG_k(t,r)f(r,\cdot)(x)\to D_tG(t,r)f(r,\cdot)(x)$ as $k\to+\infty$ for any $(t,r,x)\in(\widetilde{E}(T)\cap[(\mathcal{F}\cap E_0)\times[0,T]])\times\mathbb{R}^n$. Here $\mathcal{F}$ denotes the set of all $t\in[0,T]$ such that $Q_0^{(k)}$, $B_0^{(k)}$, and $c^{(k)}$ converge to Q_0, B_0, and c, respectively. Hence $M(f)$ is measurable in $[0,T]^2\times\mathbb{R}^n$. Then, reasoning as in the case of $L(f)$, it can be easily proved that $M(f)$ is measurable for any measurable function f such that f/p is bounded.

By the same technique, it can be proved that the function $(t,r)\to M(f)(t,r,x)$ is measurable for any $x\in\mathbb{R}^n$. Next, by virtue of the dominated convergence theorem, it can be easily proved that for any t and any $x\in\mathbb{R}^n$, the function $M(f)(t,\cdot,x)$ is measurable in $[0,T]$. Then, reasoning as in the proof of the measurability of the function $L(f)$ it can be easily shown that for any $(r,x)\in[0,T]\times\mathbb{R}^n$ the function $M(f)(\cdot,r,x)$ is measurable in $[0,T]$.

To conclude we consider the function $N(f)$. By the previous results, we already know that the function $r\to D_tG(t,r)f(r,\cdot)(x)$ is measurable in $[0,T]$ for any $(t,x)\in E\times\mathbb{R}^n$. Moreover, thanks to Theorem 3.5 and Lemma 3.7, we deduce that there exists a positive constant D such that $\|D_tG(t,r)f(r,\cdot)(x)\|\le Dp(x)|t-r|^{1-\theta/2}$ a.e. in $[0,t]$. Consequently, $N(f)$ is well defined. Then we observe that $M(f)$ is integrable in $[0,T]^2\times K$ for any compact set $K\subset\mathbb{R}^n$. Thanks to the Fubini–Tonelli theorem, we easily deduce that $N(f)$ is a measurable function in $[0,T]\times\mathbb{R}^n$. □

LEMMA A.2. *Suppose that*

(i) $u\in\mathrm{Lip}\,([0,T]\times\mathbb{R}^n)$ *with support contained in a compact set* $[0,T]\times K\subset[0,T]\times\mathbb{R}^n$;

(ii) u *is twice continuously differentiable with respect to the variable* x *in* $[0,T]\times\mathbb{R}^n$;

(iii) u *is differentiable with respect to the variable* t *for any* $(t,x)\in F\subset[0,T]\times\mathbb{R}^n$ *and* $D_tu(t,x)=\mathcal{A}(t,x)u(t,x)$ *for any* $(t,x)\in F$, *where* F *is a measurable set such that its complement in* $[0,T]\times\mathbb{R}^n$ *is negligible and for any* $x\in\mathbb{R}^n$, *the set* $F(x)=\{t\in[0,T]:\ (t,x)\in F\}$ *is measurable in* $[0,T]$ *with measure* T.

Then there exists a measurable set $A\subset[0,T]$ *with measure* T *such that for any* $(t,\xi)\in A\times\mathbb{R}^n$, *we have*

$$D_t\widehat{u}(t,\xi)=\left(-\sum_{i,j=1}^n q_{i,j}(t)\xi_i\xi_j+\mathrm{i}\sum_{j=1}^n b_j(t)\xi_j+c(t)\right)\widehat{u}(t,\xi),$$

where $\widehat{u}(t,\cdot)$ *denotes the Fourier transform of the function* $u(t,\cdot)$.

Proof. By assumption, F^c is negligible. Therefore, there exists a measurable set A with measure T such that $\mathbb{R}^n\backslash F(t)$ is negligible for any $t\in A$. Here $F(t)=\{x\in\mathbb{R}^n:\ (t,x)\in F\}$. Hence for any $t_0\in A$, we have

$$\begin{aligned}&(t-t_0)^{-1}\int_{\mathbb{R}^n}\big[u(t,x)-u(t_0,x)\big]\exp(-\mathrm{i}\,x\cdot\xi)dx\\&\quad=(t-t_0)^{-1}\int_{F(t_0)}\big[u(t,x)-u(t_0,x)\big]\exp(-\mathrm{i}\,x\cdot\xi)dx.\end{aligned}$$

Next we observe that $u(\cdot,x)$ is differentiable with respect to the variable t in t_0 for any $x \in F(t_0)$ and

$$|u(t,x)-u(t_0,x)| \le [u]_{\mathrm{Lip}\,(\mathbb{R}^n)}\chi_K|t-t_0|.$$

Applying the dominated convergence theorem, we deduce that the Fourier transform of u is differentiable with respect to the variable t at each point $(t,x) \in A \times \mathbb{R}^n$ and

$$D_t\widehat{u}(t,\xi) = \int_{F(t)} D_t u(t,x)\exp(-\mathrm{i}\,x\cdot\xi)dx.$$

By assumption,

$$D_t u(t,x) = \sum_{i,j=1}^n q_{i,j}(t)D^2_{i,j}u(t,x) + \sum_{j=1}^n b_j(t)D_j u(t,x) + c(t)u(t,x) \qquad \forall (t,x)\in F.$$

In particular, the previous relationship holds for any $t \in A$ and any $x \in F(t)$. Hence

$$\begin{aligned} D_t\widehat{u}(t,\xi) &= \sum_{i,j=1}^n q_{i,j}(t)\int_{\mathbb{R}^n} D^2_{i,j}u(t,x)\exp(-\mathrm{i}\,x\cdot\xi)dx \\ &\quad + \sum_{j=1}^n b_j(t)\int_{\mathbb{R}^n} D_j u(t,x)\exp(-\mathrm{i}\,x\cdot\xi)dx + c(t)\int_{\mathbb{R}^n} u(t,x)\exp(-\mathrm{i}\,x\cdot\xi)dx \\ &= \left(-\sum_{i,j=1}^n q_{i,j}(t)\xi_i\xi_j + \mathrm{i}\sum_{j=1}^n b_j(t)\xi_j + c(t)\right)\widehat{u}(t,\xi) \quad \forall (t,\xi)\in A\times\mathbb{R}^n. \end{aligned}$$

□

LEMMA A.3. *Suppose that f is a measurable function belonging to $B([0,T];C(\mathbb{R}^n))$. Then for any $r \in [0,T]$ and any $x \in \mathbb{R}^n$, the functions $f(r,\cdot)$ and $f(\cdot,x)$ are measurable in $\mathbb{R}^n$ and $[0,T]$, respectively.*

Proof. We begin by observing that, trivially, for any $r \in [0,T]$, the function $f(r,\cdot)$ is measurable in $\mathbb{R}^n$. Next we consider the function $f(\cdot,x)$. By Fubini's theorem, there exists a measurable set $\mathcal{D} \subset \mathbb{R}^n$ such that its complement is negligible and the function $f(\cdot,x)$ is measurable in $[0,T]$ for any $x \in \mathcal{D}$. Then with any $x \in \mathbb{R}^n$ we associate a sequence $\{x_n\}_{n\in\mathbb{N}} \subset \mathcal{D}$ converging to x as $n \to +\infty$. Since $f(r,\cdot) \in C(\mathbb{R}^n)$ for any $r \in [0,T]$, we immediately deduce that $f(\cdot,x)$ is the pointwise limit of the sequence of measurable functions $f(\cdot,x_n)$. Consequently, $f(\cdot,x)$ is measurable in $[0,T]$. □

COROLLARY A.4. *Let p be either the polynomial or the exponential weight function and f a measurable function belonging to $B([0,T];C^\theta_p(\mathbb{R}^n))$. Then for any $r \in [0,T]$ and any $x \in \mathbb{R}^n$, the functions $f(r,\cdot)$ and $f(\cdot,x)$ are measurable in $\mathbb{R}^n$ and $[0,T]$, respectively.*

Proof. It suffices to apply Lemma A.3 to the function f/p. □

LEMMA A.5. *Suppose that p is either the polynomial or the exponential weight function. Then for any function $f : [0,T]\times\mathbb{R}^n \to \mathbb{R}$ such that f/p is a bounded and Lipschitz continuous, the function $t \to \|f\|_{B([0,t];C_p(\mathbb{R}^n))}$ is Lipschitz continuous in $[0,T]$.*

Proof. It suffices to observe that for any t_1, $t_2 \in [0,T]$, we have

$$\left|\|f\|_{B([0,t_2];C_p(\mathbb{R}^n))} - \|f\|_{B([0,t_1];C_p(\mathbb{R}^n))}\right| \le [f/p]_{\mathrm{Lip}\,([0,T]\times\mathbb{R}^n)}|t_2-t_1|.$$

□

Acknowledgments. The author wishes to thank the referees for their useful suggestions and comments that have improved this paper.

REFERENCES

[1] M. Bramanti and M. Cerutti, $W_p^{1,2}$ *solvability for the Cauchy-Dirichlet problem for parabolic equations with VMO coefficients*, Comm. Partial Differential Equations, 18 (1993), pp. 1735-1763.

[2] A. Brandt, *Interior Schauder estimates for parabolic differential-(or difference-)equations via the maximum principle*, Israel J. Math., 7 (1969), pp. 254–262.

[3] F. Chiarenza, M. Frasca, and P. Longo, *Interior $W^{2,p}$ estimates for non divergence elliptic equations with discontinuous coefficients*, Ricerche Mat., 40 (1991), pp. 149–168.

[4] S. D. Eidel'man, *On fundamental solutions of parabolic systems* II, Mat. Sb. (N.S.), 53 (95) (1961), pp. 73–136 (in Russian); Amer. Math. Soc. Transl., 41 (1964), pp. 49–120 (in English).

[5] A. Friedman, *Partial Differential Equations of Parabolic Type*, Prentice–Hall, Englewood Cliffs, NJ, 1964; reprinted by Krieger, Malabar, FL, 1983.

[6] B. Knerr, *Parabolic interior Schauder estimates by the maximum principle*, Arch. Rational Mech. Anal., 75 (1980), pp. 51–58.

[7] S. N. Kruzhkov, A. Castro, and M. Lopes, *Mayoraciones de Schauder y teorema de existencia de las soluciones del problema de Cauchy para ecuaciones parabolicas lineales y no lineales* I, Cienc. Mat. (Havana), 1 (1980), pp. 55–76.

[8] S. N. Kruzhkov, A. Castro, and M. Lopes, *Mayoraciones de Schauder y teorema de existencia de las soluciones del problema de Cauchy para ecuaciones parabolicas lineales y no lineales* II, Cienc. Mat. (Havana), 3 (1982), pp. 37–56.

[9] S. N. Kruzhkov, A. Castro, and M. Lopes, *Schauder type estimates and theorems on the existence of the solution of fundamental problems for linear and nonlinear parabolic equations*, Dokl. Akad. Nauk SSSR, 220 (1975), pp. 277–280 (in Russian); Soviet Math. Dokl., 16 (1975), pp. 60–64 (in English).

[10] G. Lieberman, *Intermediate Schauder theory for second order parabolic equations* IV: *Time irregularity and regularity*, Differential Integral Equations, 5 (1992), pp. 1219–1236.

[11] L. Lorenzi, *Schauder estimates for the Ornstein-Uhlenbeck semigroup in spaces with polynomial or exponential growth*, Dynam. Systems Appl., 9 (2000), pp. 199–219.

[12] A. Lunardi, *Analytic Semigroups and Optimal Regularity in Parabolic Problems*, Birkhauser-Verlag, Basel, 1995.

[13] A. Lunardi, *An interpolation method to characterize domains of generators of semigroups*, Semigroup Forum, 53 (1996), pp. 321–329.

[14] C. P. Mawata, *Schauder estimates and existence theory for entire solutions of linear parabolic equations*, Differential Integral Equations, 2 (1989), pp. 251–274.

[15] H. Triebel, *Interpolation Theory, Functional Spaces, Differential Operators*, North-Holland, Amsterdam, 1978.

SIAM J. MATH. ANAL.
Vol. 32, No. 3, pp. 616–636

OPTIMAL APPROXIMATIONS OF TRANSPORT EQUATIONS BY PARTICLE AND PSEUDOPARTICLE METHODS*

ALBERT COHEN† AND BENOIT PERTHAME†

Abstract. The convergence rate of particle methods for solving linear transport equations is revisited. Denoting h the initial discretization parameter, we prove a quasi-optimal rate of convergence like $h^{s-\varepsilon}$ for all $\varepsilon > 0$ for an initial data in the Sobolev space $W^{s,p}$ when choosing appropriate initial integration rules and general convolution. As it is well known, this suboptimality is due to the form and width of the convolution kernel. In particular, it can be fixed by computing an additional quantity, the cell deformation. Then one can restore the optimal rate of convergence, up to the first order ($s = 1$), while keeping the built-in conservative aspect. To avoid these additional computations and move to higher order optimality, another strategy is introduced and analyzed. It is based on a discretization of the solution at initial time by local averages but differs from the usual particle methods: the local averages are viewed as point values of an approximation of the solution, and the regularization of the solution at time $t > 0$ is performed by interpolation rather than convolution. This strategy allows us to recover optimal error estimates in L^p or Sobolev norms (up to any prescribed order).

Key words. transport equations, particle methods, vortex methods

AMS subject classifications. 65M12, 65L05, 76M25

PII. S0036141099350353

1. Introduction.

Particle methods are of common use for the numerical simulation of transport equations. These methods typically operate in three steps:

(i) The initial value $u_0(x) := u(x, t = 0)$ is approximated in the distribution sense by a linear combination $v_0(x) := \sum_k \alpha_k \delta_{x_k}$ of Dirac masses, with weights α_k that represent the integral of u_0 in a neighborhood the point x_k.

(ii) One follows the time evolution of the *p*article solution that corresponds to the initial measure v_0. Due to the form of the equation, this exact solution (in the distribution sense) can be written $v(x,t) = \sum_k \alpha_k(t)\delta_{x_k(t)}$. The evolution of the weights $\alpha_k(t)$ and locations $x_k(t)$ of the particles is described by ordinary differential equations that can be simulated by high order numerical techniques.

(iii) In order to recover a proper approximation of the solution $u(x,t)$ at some time $t > 0$, one needs to regularize the particle solution $v(x,t)$. Such a regularization is usually performed by a convolution product with a so-called "cut-off function" φ after a proper scaling $\varphi_\varepsilon(x) := \varepsilon^{-d}\varphi(x/\varepsilon)$ that takes into account the initial tightness of the particle discretization.

These methods have been widely used to solve conservative nonlinear transport equations: see [11] and [7] for recent applications to incompressible fluid dynamics (more precisely, vortex methods), and [10] for applications to kinetic equations (more precisely, Vlasov equations). In a linear context, they have been analyzed in [14] (see also [7] and [13] for a general introduction to these methods and their numerical analysis). We recall here the main features in the specific setting of the model

*Received by the editors January 4, 1999; accepted for publication (in revised form) March 3, 2000; published electronically October 20, 2000.
http://www.siam.org/journals/sima/32-3/35035.html
†Université Pierre et Marie Curie et CNRS UMR 7598, Laboratoire d'Analyse Numérique, BC187, 4, pl. Jussieu, 75252 Paris Cédex 05, France (cohen@ann.jussieu.fr, perthame@ann.jussieu.fr).

transport problem

$$\frac{\partial u}{\partial t} + \sum_{i=1}^{d} \frac{\partial}{\partial x_i}(a_i u) + a_0 u = f, \quad x \in \mathbb{R}^d, \ 0 \le t \le T, \tag{1.1}$$

with some initial data $u_0(x) := u(x, 0)$. In order to simplify the analysis, we assume here that the data f and the coefficients a_i are C^∞ functions in space and time, and that a_i and its derivatives are uniformly bounded on $\mathbb{R}^d \times [0, T]$, $1 \le i, j \le d$. These assumptions ensure that the evolution operator

$$E_t : \ u(0) \mapsto u(t) := u(\cdot, t) \tag{1.2}$$

is well defined and bounded in all Sobolev spaces $W^{s,p}$, $s \ge 0$, $1 \le p \le \infty$ (see [14]). It can be built through the method of characteristics. One can define a family of C^∞ diffeomorphisms Φ_t for $0 \le t \le T$, by $\Phi_t(x) = z(t)$ where z is the solution of

$$\frac{\partial z_i}{\partial t} = a_i(z), \quad i = 1, \ldots, d, \quad z(0) = x. \tag{1.3}$$

When the initial data u_0 is smooth enough, say, C^1, the classical solution of (1.1) exists and is built as follows: for all $x \in \mathbb{R}^d$, the value $v_x(t) = u(\Phi_t(x), t)$ is given by an ordinary differential equation

$$\frac{d}{dt} v_x(t) = f(\Phi_t(x), t) - \tilde{a}_0(\Phi_t(x), t) v_x(t), \quad v_x(0) = u_0(x), \tag{1.4}$$

with $\tilde{a}_0 := a_0 + \mathrm{div} a$. In particular the value is preserved on the characteristics curves $\Phi_t(x)$ if $f = \tilde{a}_0 = 0$.

Particle methods allow us to treat solutions that are not necessarily continuous. They are particularly well adapted to the conservative form of (1.1) as shown by the following example: given an initial discretization parameter h, one defines a weak approximation of u_0 by the measure

$$v_h(0) := \sum_{k \in \mathbb{Z}^d} \alpha_k \delta(\cdot - kh), \tag{1.5}$$

where δ_x denotes the Dirac function at some point x and

$$\alpha_k := \int_{kh + Q_h} u_0(x) dx, \quad Q_h := \left[-\frac{h}{2}, \frac{h}{2} \right]^d. \tag{1.6}$$

If $f = 0$, the solution of (1.1) in the distribution sense corresponding to the initial data (1.5) is exactly given at time t by

$$v_h(t) = \sum_{k \in \mathbb{Z}^d} \alpha_k(t) \delta(\cdot - x_k(t)), \tag{1.7}$$

where $x_k(t) = \Phi_t(kh)$ and the weights $\alpha_k(t)$ are solutions of the ordinary differential equations

$$\frac{d}{dt} \alpha_k(t) + a_0(x_k(t), t) \alpha_k(t) = 0, \quad \alpha_k(0) = \alpha_k. \tag{1.8}$$

The value of $\alpha_{k,t}$ is an approximation of the integral of the real solution u_t over the transported domain $\Phi_t(kh+Q_h)$. In particular, this approximation is consistent with the conservation of this quantity when $a_0 = 0$.

When the source term is nonzero, it needs to be approximated in a proper way: a possibility is to take

$$f_h = \sum_{k\in\mathbb{Z}^d} \beta_k(t)\delta(x - x_k(t)), \tag{1.9}$$

where $\beta_{k,t}$ takes account of the source f over $\Phi_t(kh + Q_h)$. Since the shape of this transported domain is not exactly known in practice, one usually uses a quadrature formula of the type

$$\beta_k(t) := w_k(t)f(x_k(t), t), \quad w_k(t) = h^d|\det(D\Phi_t)(kh)|. \tag{1.10}$$

The particle solution is then again given by (1.7) with weights satisfying the ordinary differential equations

$$\frac{d}{dt}\alpha_k(t) + a_0(x_k(t), t)\alpha_k(t) = \beta_k(t), \quad \alpha_k(0) = \alpha_k. \tag{1.11}$$

The regularization by convolution with $\varphi_\varepsilon := \varepsilon^{-d}\varphi(\cdot/\varepsilon)$ yields an approximation $u_{h,\varepsilon}(t)$ of the real solution $u(\cdot, t)$ at time $t \in]0, T]$, with the expression

$$u_{h,\varepsilon}(t) := v_h(t) * \varphi_\varepsilon = \sum_{k\in\mathbb{Z}^d} \alpha_k(t)\varphi_\varepsilon(\cdot - x_k(t)). \tag{1.12}$$

For a given $h > 0$, the value of $\varepsilon > 0$ needs to be chosen in such a way that the error $\|u(t)-u_{h,\varepsilon}(t)\|$ in some prescribed norm—or an estimation of this error—is optimized.

Intuitively, this optimal choice solves a compromise. On the one hand, if ε is small in comparison to the minimal distance between the points $x_k(t)$, the approximate solution defined by (1.12) will vanish away from these points and is thus irrelevant. On the other hand, if ε is too large, the approximate solution will contain only low frequencies and will necessarily lack accuracy.

This compromise is expressed in the classical error estimate for such methods [14]. If φ is a $W^{m,1}$ function such that $\int \varphi = 1$, $\int |x|^r|\varphi(x)|dx < \infty$ and

$$\int x_1^{k_1} \cdots x_d^{k_d}\varphi(x_1, \ldots, x_d) = 0, \quad |k| := k_1 + \cdots + k_d \le r - 1 \tag{1.13}$$

for some prescribed $m > 0$ and integer $r > 0$, then the following error estimate can be proved to hold: there is a constant $C(t)$ depending only on the time t such that for all $0 < h \le \varepsilon \le 1$,

$$\|u(t) - u_{h,\varepsilon}(t)\|_{L^p} \le C(t)\Big(\varepsilon^r|u_0|_{W^{r,p}} + (h/\varepsilon)^m\|u_0\|_{W^{m,p}}\Big). \tag{1.14}$$

The optimization of (1.14) obliges to choose $\varepsilon \sim h^{m/(m+r)}$, yielding a suboptimal estimate in $h^{mr/(m+r)}$, depending strongly on the choice of the function φ (through the smoothness m and the number r of vanishing moments). With the best possible choice of φ and the parameters, for a given initial regularity $W^{s,p}$ (i.e., with φ such that $s = m = r\}$), one thus obtains only the error estimate $h^{s/2}\|u_0\|_{W^{s,p}}$. However, since the particle method is based on a discretization of step size h, an optimal result should yield an estimate in $Ch^s\|u_0\|_{W^{s,p}}$.

The explanation of this lack of accuracy is simple. The necessity to choose ε substantially larger than h is somehow related to the fact that for $t > 0$ the grid of particle $(x_k(t))_{k\in\mathbb{Z}^d}$ is no longer uniform. In contrast, one can easily design a function φ such that the optimal estimate

$$\|u(t) - u_{h,h}(t)\|_{L^p} \leq C(t)h^s\|u_0\|_{W^{s,p}} \tag{1.15}$$

holds at time $t = 0$. To do so, it is sufficient to impose that the operator

$$f \mapsto Pf := \sum_{k\in\mathbb{Z}^d} \langle f, \chi_{Q_1}(\cdot - k)\rangle\varphi(\cdot - k) \tag{1.16}$$

reproduces polynomials up to order $[s]$ and use classical approximation theory arguments as in [2] or [15]. Concrete examples of functions φ satisfying such properties are provided in particular by the biorthogonal wavelet theory developed in [6] (see also Chapter II in [5]). The suboptimality of (1.14) reveals a weakness of the convolution method for regularization: in terms of approximation theory, convolution is better adapted to translation invariant data, and optimal error rates are lost out of this setting. On the other hand, as h goes to zero, the spacing between the particles at time t is still in $\mathcal{O}(h)$ (with constants depending on t), so that it is reasonable to search for an alternate method that yields the optimal error estimate in $h^s\|u_0\|_{W^{s,p}}$.

In the present paper, we propose to go further in the search for accuracy. A first possibility, which yields a quasi-optimal method in $h^{s-\beta}$, $\beta > 0$, is to play with the initial integration rule. We choose a better formula than (1.6), replacing the indicator function by a function $\tilde{\varphi}$. This yields

$$\alpha_k = \int u_0(x)\tilde{\varphi}(x - kh)dx.$$

We first show that, with appropriate functions $\tilde{\varphi}$, one can already obtain this quasi-optimal error estimate. This result is already interesting because it proves that particle methods are potentially better than finite volumes methods (another conservative method) which only converge with the rate $h^{1/2}$ for BV data. A second method consists of using a better reconstruction of $u_{h,\varepsilon}$. We show that it is possible to adapt locally the width of the convolution kernel with the help of new differential equations which describe the deformation of the initial grid. Then an optimal rate can be restored, keeping the natural conservation property of particle methods, but this method is limited to first order estimates.

To go further we propose a third method to restore the optimal error estimate at any order and without solving new differential equations. It is based on a different regularization technique. The main idea underlying this technique is to approximate the initial value by a smooth function $v_{0,h}$ whose point values on the grid $h\mathbb{Z}^d$ are given by local averages of u_0 and to follow the evolution of this function on the transported grid using the ordinary differential equation (1.4). At time $t > 0$, convolution is then replaced by an interpolation process that requires a particular design in order to obtain the optimal result. We refer to this strategy as a "pseudoparticle" method, since formally it also amounts to solving ordinary differential equations for each trajectory related to an initial grid point, but the solutions no longer describe the evolution of a combination of Dirac masses and thus the local conservation of mass is lost (it can nevertheless be restored by a correction that does not change the optimal order of accuracy). It should pointed out that interpolation techniques have already been

considered in [1] for the purpose of improving the accuracy of the vortex methods at large time.

The outline of the paper is the following. We first present in section 2 the quasi-optimal method based on appropriately choosing the initial weights. In section 3, we prove the optimal estimate based on convolution with a locally varying kernel. A very simple one-dimensional example of the pseudoparticle strategy (and a proof of its optimality) is described in section 4. We then describe in section 5 the pseudoparticle method, in particular the discretization of the initial value u_0 and the interpolation process at time $t > 0$. Finally, we derive in section 6 the optimal error estimates for this method. Throughout the paper, $C(t)$ will always denote a constant which is independent of the various parameters (initial value, mesh size h) but may vary with t.

2. A quasi-optimal estimate with improved initial quadrature. We consider the particle methods described in the introduction with $f = 0$ for the sake of simplicity. In this section, we show a quasi-optimal a priori L^p-estimates, in $C_\beta h^{s-\beta}$ for all $\beta > 0$, when the initial data u_0 has $W^{s,p}$ smoothness. It is obtained by improving the initial weights through a particular initial "quadrature" formula based on an appropriate averaging function $\tilde{\varphi}$. Notice that at initial time $t = 0$, the optimal approximation rate h^s can be obtained by choosing the pair averaging-convolution $(\tilde{\varphi}, \varphi)$ with particular relations; see (1.16) and sections 4 and 5. However, this optimality is lost in general for time $t > 0$ due to the particle dynamic and considerations on $\tilde{\varphi}$ are enough for the present quasi-optimal result.

More specifically, for our first construction we consider a compactly supported function $\tilde{\varphi} \in C^0(\mathbb{R}^d)$ which satisfies the conditions

$$\sum_{k\in\mathbb{Z}^d} \tilde{\varphi}(. - k) = 1, \tag{2.1}$$

$$\sum_{k\in\mathbb{Z}^d} k_1^{p_1} \cdots k_d^{p_d} \tilde{\varphi}(y - k) = y_1^{p_1} \cdots y_d^{p_d}, \quad |p| \le m - 1. \tag{2.2}$$

In the particle method described in the introduction, we use the initial weights (again with the notation $\tilde{\varphi}_h(y) = \tilde{\varphi}(h^{-1}y)$)

$$\alpha_k(0) = \int_{\mathbb{R}^d} u_0(y)\tilde{\varphi}_h(y - kh)dy. \tag{2.3}$$

We also introduce a compactly supported continuous cut-off $\varphi \in W^{m,1}$ such that

$$\int_{\mathbb{R}^d} \varphi = 1, \tag{2.4}$$

$$\int_{\mathbb{R}^d} x_1^{k_1} \cdots x_d^{k_d} \varphi(x_1, \ldots, x_d) = 0, \ \ |k| \le r - 1, \tag{2.5}$$

and $\int_{\mathbb{R}^d} |x|^r |\varphi(x)|dx < \infty$. Then, defining the particle solution as in the introduction, we have the following result.

THEOREM 2.1. *We assume* (2.1), (2.2) for some $m > \frac{d}{p'}$ and $1 \le p \le \infty$, (2.4), and (2.5). Then there is a constant $C(m, t)$ which also depends on r, φ, and $\tilde{\varphi}$ such that for all $0 < h \le \varepsilon$,

$$\|u(t) - u_{h,\varepsilon}(t)\|_{L^p} \le C(t, m)\Big(\varepsilon^r \|u_0\|_{W^{r,p}} + (h/\varepsilon)^m \|u_0\|_{L^p}\Big). \tag{2.6}$$

Proof. As in the classical analysis of particle methods (see [14], [13], or [7]), we set $u_\varepsilon = u * \varphi_\varepsilon$ and we have to estimate

$$\|u(t) - u_{h,\varepsilon}(t)\|_{L^p} \le \|u(t) - u_\varepsilon(t)\|_{L^p} + \|u_\varepsilon(t) - u_{h,\varepsilon}(t)\|_{L^p}. \tag{2.7}$$

The first term in the right-hand side of the above inequality is classically upper bounded by $C(s,\varphi)\varepsilon^r \|u_0\|_{W^{r,p}}$ thanks to the assumptions (2.4), (2.5).

The second term is first treated as usual by an inverse estimate which yields

$$\begin{aligned} \|u_\varepsilon(t) - u_{h,\varepsilon}(t)\|_{L^p} &= \|\varphi_\varepsilon * (u(t) - v_h(t))\|_{L^p} \\ &\le C\varepsilon^{-m} \|u(t) - v_h(t)\|_{W^{-m,p}} \\ &\le C(t)\varepsilon^{-m} \|u(0) - v_h(0)\|_{W^{-m,p}}, \end{aligned}$$

where the last inequality comes from the stability of the evolution operator in $W^{-m,p}$. Introducing next the operators

$$u \mapsto P_h u := \sum_{k\in\mathbb{Z}^d} \langle u, \tilde\varphi_h(y - kh)\rangle \delta_{kh} \tag{2.8}$$

and

$$u \mapsto P_h^* u := \sum_{k\in\mathbb{Z}^d} u(kh)\tilde\varphi_h(y - kh), \tag{2.9}$$

we see that

$$\begin{aligned} \|u(0) - v_h(0)\|_{W^{-m,p}} &= \sup_{\|w\|_{W^{m,p'}}=1} |\langle w, u(0) - P_h u(0)\rangle| \\ &= \sup_{\|w\|_{W^{m,p'}}=1} |\langle w - P_h^* w, u(0)\rangle| \\ &\le C\|u(0)\|_{L^p} \sup_{\|w\|_{W^{m,p'}}=1} \|w - P_h^* w\|_{L^{p'}}, \end{aligned}$$

with $1/p' + 1/p = 1$. We finally remark that the assumptions (2.1) and (2.2) ensure that the interpolation operator P^* reproduces polynomials up to the degree $m-1$. Thus classical approximation theory argument yield the direct estimate

$$\|w - P_h^* w\|_{L^{p'}} \le Ch^m \|w\|_{W^{m,p'}}, \tag{2.10}$$

provided that $m > d/p'$ so that P_h^* can be applied to the functions in $W^{m,p'}$ which are then continuous. This allows us to derive the estimate in $(h/\varepsilon)^m \|u_0\|_{L^p}$ for the second term of (2.7). ∎

Remark 2.1. In particular, if $s \le r$, the optimal choice of the cut-off parameter $\varepsilon = h^{m/(m+s)}$ gives the convergence rate $h^{sm/(m+s)}$. As m tends to infinity, this rate tends to the optimal value h^s. But, whatever is the choice of $\tilde\varphi$, we should expect that the constant $C(t,m)$ tends to infinity with m.

Remark 2.2. Our method is in essence close to the method introduced in [3], which performs regularization of the initial condition before particle discretization. However, in [3], this regularization is performed at scale ε, resulting in more costful computations for the initial weights.

Remark 2.3. We can give examples of functions $\tilde\varphi$. The function Q_1 only satisfies (1.1) and thus the result holds with $m = 1$. For the hat function, see (4.1) below, one readily checks that the conditions (2.2) hold with $m = 2$. More generally, one can build examples at any order m in various manners. A possibility is to take the so-called

Deslaurier–Dubuc interpolatory scaling functions ϕ_m which are the autocorrelation $\varphi_m * \varphi_m(-\cdot)$ of the Daubechies orthonormal scaling function with support in $[0, 2m-1]$ (see Chapter II in [5]). In this case, (2.2) holds with order $2m$, the case $m = 1$ corresponding to the hat function $\phi_1 = \chi_{[0,1]} * \chi_{[-1,0]}$. Another possibility is to use cardinal splines of odd degree $2p + 1$, but these are nonlocal functions for $p \geq 1$. A last simple possibility is to replace the functions $\tilde{\varphi}_h(\cdot - kh)$ by the nodal basis functions of quadrilateral Lagrange finite elements of degree $m - 1$: these bases are not obtained from the shifts of a single function (except for $m = 2$), but they still yield an interpolation operator P_h^* with good approximation properties, which is the main tool needed in the above proof.

Remark 2.4. In the case of L^1 estimates, one can use BV smoothness in place of $W^{1,1}$ to derive the same rate, which allows discontinuities even in one dimension for the solution and is useful for applications.

3. A first order estimate with local convolutions. In this section we show that it is possible to restore the right order of convergence for particles methods in using a local convolution operator. Then the idea is to use a convolution with the indicator function of the local cell obtained in deforming the initial cell by the linearized flow—this depends of the point $x_k(t)$ under consideration. The advantage of this method is to keep the built-in local conservation of "mass" useful for many applications. The drawback is that, in order to adapt locally the convolution, we need to follow additional quantities (the deformation by the flow of the tangent space). This idea has already been used in the vortex case, in order to reduce the constant in the error estimate, by T.Y. Hou [12].

In order to simplify the analysis we limit ourselves to first order approximations in L^1 norms and we consider only the simplest case of initial weights given by (1.6). We introduce at time t the parallelepipedic cell $Q_{k,h}(t)$ obtained in deforming the initial cells Q_h with the linearized flow around the trajectory $x_k(t)$. In other words, we consider the linear operators (here the matrix Da is $\frac{\partial a_i}{\partial x_j}$)

$$\frac{d\mathcal{L}_{k,h}(t)}{dt} = Da\Big(x_k(t)\Big) \cdot \mathcal{L}_{k,h}(t), \tag{3.1}$$

$$\mathcal{L}_{k,h}(0) = Id. \tag{3.2}$$

And $Q_{k,h}(t) = \mathcal{L}_{k,h}(t) \cdot Q_h$. Next we consider the reconstruction

$$\tilde{u}_h(t,x) = \sum_{k \in \mathbb{Z}^d} \alpha_k(t) Vol\Big(Q_{k,h}(t)\Big)^{-1} \chi_{Q_{k,h}(t)}(x - x_k(t)), \tag{3.3}$$

where χ_E denotes the indicator function of the set E. Then we have the following theorem.

THEOREM 3.1. *We make the assumptions of the introduction for* (1.1) *and the above construction. Then there is a constant such that*

$$\|u(t) - \tilde{u}_h(t)\|_{L^1} \leq C(t)h \tag{3.4}$$

for all $u_0 \in BV(\mathbb{R}^d)$.

Remark 3.1. In the conservative case, $a_0 = 0$, we have

$$\int_{\mathbb{R}^d} \tilde{u}_h(t,x)dx = \sum_{k\in\mathbb{Z}^d} \alpha_k(t) Vol\Big(Q_{k,h}(t)\Big)^{-1} Vol\Big(Q_{h,k}(t)\Big)$$

$$= \sum_{k\in\mathbb{Z}^d} \alpha_k(0) = \int_{\mathbb{R}^d} u^0(x)dx,$$

and the method is indeed conservative.

Remark 3.2. It is possible to reach a second order of accuracy, for $u_0 \in W^{2,1}$, but this requires us to follow in time the second deformations of the cells. Another issue is to reconstruct the cells from the only points $x_k(t)$ since they give asymptotically a good approximation of the deformed cells. This point of view is treated in sections 4–6.

Proof. We again simplify the notations and consider only the case $a_0 = 0$. To begin with, we interpret our reconstruction of $\tilde{u}_h$ as a local convolution introducing the function $\varphi(t,y,z)$ as

$$\begin{aligned} \tilde{u}_h(t,x) &= \sum_{\mathbb{Z}^d} \alpha_h(t) Vol\Big(Q_{k,h}(t)\Big)^{-1} \chi_{Q_{k,h}(t)}(x - x_k(t)) \\ &= \int_{\mathbb{R}^d} v_h(t,y)\varphi(t,y,x-y)dy, \end{aligned} \tag{3.5}$$

where

$$\varphi(t,y,z) = Vol\Big(Q_{Y(0),h}(t)\Big)^{-1} \chi_{Q_{Y(0),h}(t)}(z),$$

the value of $Y(0)$ as a function of y being given through the characteristics

$$\frac{dY(s)}{ds} = a\Big(Y(s)\Big), \quad Y(t) = y.$$

This function $\varphi(t,y,z)$ satisfies the equation

$$\frac{\partial\varphi}{\partial t} + a(y).\nabla_y\varphi + Da(y).\nabla_z(z\varphi) = 0, \tag{3.6}$$

$$\varphi(0,x,z) = h^{-d}\chi_{Q_h}(z). \tag{3.7}$$

To prove this, we introduce the coupled trajectories

$$\frac{dY(t)}{dt} = a\Big(Y(t)\Big), \quad Y(0) = y,$$

$$\frac{dZ(t)}{dt} = Da\Big(Z(t)\Big).Z(t), \quad Z(0) = z.$$

Then the function $\psi(t,Y(t),Z(t)) = \chi_{Q_{Y(0),h}(t)}\Big(Z(t)\Big)$ satisfies

$$\psi(t,Y(t),Z(t)) = \chi_{Q_h}(z) = \psi(0,y,z).$$

Indeed, with a notation similar to (3.1), $Z(t) = \mathcal{L}_y(t).z$, therefore $Z(t) \in Q_{Y(0),h}(t)$ if and only if $z \in Q_h$. Therefore we have

$$\frac{\partial\psi}{\partial t} + a(y).\nabla_y\psi + Da(y).z\ \nabla_z\psi = 0,$$

and the equation on φ follows after some algebraic manipulations of this equation.

We now come back to estimate the error. We write the equation on $\tilde{u}_h(t,x)$, thanks to (3.5), and using the notation $z = x - y$, $\varphi = \varphi(t, y, x - y)$,

$$\frac{\partial \tilde{u}_h}{\partial t} + \mathrm{div}_x\Big(a(x).\tilde{u}_h\Big)$$

$$= \int_{\mathbb{R}^d} \left[\frac{\partial v_h(t,y)}{\partial t}\varphi(t,y,x-y) + v_h(t,y)\left[\frac{\partial \varphi}{\partial t} + \mathrm{div}_x\Big(a(x)\varphi\Big)\right]\right] dy$$

$$= \int_{\mathbb{R}^d} v_h(t,y)\left[\frac{\partial \varphi}{\partial t} + a(y)(\nabla_y\varphi - \nabla_z\varphi) + \varphi \mathrm{div}\ a(x) + a(x).\nabla_z\varphi\right] dy$$

$$= \int_{\mathbb{R}^d} v_h(t,y)\left[\frac{\partial \varphi}{\partial t} + a(y)\nabla_y\varphi + \varphi \mathrm{div}\ a(x) + [a(x) - a(y)].\nabla_z\varphi\right] dy$$

$$= \int_{\mathbb{R}^d} v_h(t,y)\left[\frac{\partial \varphi}{\partial t} + a(y)\nabla_y\varphi + \varphi[\mathrm{div}\ a(y) + O(z)]\right.$$

$$\left. + [Da(y).z + O(z^2)].\nabla_z\varphi\right] dy$$

$$= \int_{\mathbb{R}^d} v_h(t,y)\left[\frac{\partial \varphi}{\partial t} + a(y)\nabla_y\varphi + Da(y).\nabla_z(z.\varphi) + O(z^2)D_z\varphi + O(z)\varphi\right] dy$$

$$= \int_{\mathbb{R}^d} v_h(t,y)[O(z^2)D_z\varphi + O(z)\varphi] dy.$$

Here we have used (3.6). Next, φ is supported by $\{|z| \le Ch\}$. Also, the regularity of the coefficients, the special structure of (3.6), and its scaling in z/h show that $\|D_z\varphi(t,y,z)\|_{L^1_z}$ is uniformly bounded in c/h and thus we obtain from the above equalities

$$\frac{\partial \tilde{u}_h}{\partial t} + \mathrm{div}_x\Big(a(x).\tilde{u}_h\Big) = hR(t,x),$$

with R uniformly bounded in $L^\infty((0,T); L^1(\mathbb{R}^d))$ for all $T > 0$. Therefore, by a comparison with the same equation for u, we obtain the desired estimate

$$\|(u - \tilde{u}_h)(t)\|_{L^1(\mathbb{R}^d)} \le \|(u - \tilde{u}_h)(0)\|_{L^1(\mathbb{R}^d)} + C(t)h \le C(t)h. \qquad \square$$

4. A simple one-dimensional example. In this section, we assume that $d = 1$, i.e., we work in one space dimension, and we consider (1.1). We shall describe the pseudoparticle method in this simple setting and prove its optimality. The main arguments (in particular the key estimates (2.4), (2.8), and (2.10) below) are only sketched here, since they are detailed in more general settings in the next section.

From the hat function

$$\varphi(x) = \max\{1 - |x|, 0\} \tag{4.1}$$

and its scaled version $\varphi_h = h^{-1}\varphi(\cdot/h)$, we can define an approximation of u_0 defined by

$$v_{0,h} = \sum_{k\in\mathbb{Z}^d} \alpha_k \varphi_h(\cdot - kh), \tag{4.2}$$

with

$$\alpha_k := \int_{(k-1/2)h}^{(k+1/2)h} u_0(x)dx. \tag{4.3}$$

Note that this approximation is exactly the convolution of the measure $v_{0,h}$ defined in (1.5) by φ_h. The operator mapping u_0 to $v_{0,h}$ is local, L^p-stable, and reproduces polynomials up to degree 1. It is well known in approximation theory [2], [15] that these properties yield the error estimate

$$\|u_0 - v_{0,h}\|_{L^p} \leq Ch^s \|u_0\|_{W^{s,p}} \tag{4.4}$$

for $0 < s \leq 2$. Note also that we have $v_{0,h}(kh) = h^{-1}\alpha_k$: the local averages over the intervals $[(k-1/2)h, (k+1/2)h]$ coincide with the point values of the approximation $v_{0,h}$.

Now define $v_h(t)$ to be the solution of (1.1) with initial value $v_{0,h}$. The values $v_h(t)$ at the points $x_k(t)$ can thus be computed since they are the solution of the ordinary differential equations (1.4) with initial value $h^{-1}\alpha_k$. Moreover, from the stability property of the evolution operator, we have

$$\|u(t) - v_h(t)\|_{L^p} \leq C(t)h^s \|u_0\|_{W^{s,p}} \tag{4.5}$$

for $0 < s \leq 2$. In order to recover an optimal approximation of $u(t)$, one can thus try to approximate $v_h(t)$ from its samples at points $x_k(t)$. Note that in one dimension, for fixed t, the sequence $(x_k(t))_{k\in\mathbb{Z}}$ is increasing and that there exists strictly positive constants $C_1(t)$ and $C_2(t)$ such that

$$C_1(t)h \leq |x_{k+1}(t) - x_k(t)| \leq C_2(t)h. \tag{4.6}$$

We can thus define a unique function $u_h(t)$ which is affine on each interval $[x_k(t), x_{k+1}(t)[$, $k \in \mathbb{Z}$ and such that $u_h(t, x_k(t)) = v_h(t, x_k(t))$. We then have the following result.

THEOREM 4.1. *For all $t \in [0, T]$, there exists a constant $C(t)$ such that*

$$\|u(t) - u_h(t)\|_{L^p} \leq C(t)h^s \|u_0\|_{W^{s,p}} \tag{4.7}$$

for all $0 < s < 1 + 1/p$.

Proof. From classical theory, the function $v_{0,h}$ satisfies an inverse estimate

$$\|v_{0,h}\|_{W^{s',p}} \leq Ch^{s-s'} \|v_{0,h}\|_{W^{s,p}} \tag{4.8}$$

for $0 \leq s < s' \leq 1 + 1/p$ (the limitation of s by above corresponds to the Sobolev regularity of the function φ which belongs to $W^{s,p}$ only for $s \leq 1 + 1/p$).

By the $W^{s,p}$-stability of the evolution operator and of the approximation operator $u_0 \mapsto v_{0,h}$, we get

$$\|v_h(t)\|_{W^{s',p}} \leq C(t)h^{s-s'} \|u_0\|_{W^{s,p}}. \tag{4.9}$$

We next use the approximation properties of the interpolation operator that defines $u_{t,h}$: from the upper inequality in (4.6) we obtain the estimate

$$\|u_h(t) - v_h(t)\|_{L^p} \leq C(t)h^{s'} \|v_h\|_{W^{s',p}} \tag{4.10}$$

for $1/p < s' < 2$ (in this case, the limitation of s' by below is necessary since $u_h(t)$ is obtained by an interpolation process which is well defined in Sobolev spaces $W^{s,p}$ only for $s > 1/p$).

Combining (4.10) and (4.9) for some s', $\max(s, 1/p) < s' < 2$, we obtain

$$\|u_h(t) - v_h(t)\|_{L^p} \leq C(t)h^s \|u_0\|_{W^{s,p}}, \tag{4.11}$$

which together with (4.5) yields the optimal estimate (4.7). ∎

Remark 4.1. The limitation $s \leq 1+1/p$ is related to the smoothness of the initial approximation, while $s \leq 2$ is related to its degree of accuracy. Both aspects should thus be considered for raising the order of the present method.

Remark 4.2. If the initial value u_0 is in $W^{s,p}$ with $s > 1/p$—and thus continuous—one can also apply the same linear interpolation process on the values of the real solution at the points $x_k(t)$. This clearly yields the same error estimate without the restriction $s \leq 1 + 1/p$ but still $s \leq 2$ due to the order of the method.

5. The pseudoparticle method. In order to raise the order of the method, we need a better approximation at the start. To do so, we consider the B-spline functions defined recursively by $B_0 = \chi_{[0,1]}$ and

$$B_n = \chi_{[0,1]} * B_{n-1} = (*)^{n+1}\chi_{[0,1]}. \tag{5.1}$$

We recall some basic properties here and refer to [1] for a general introduction: B_n is piecewise polynomial of degree n on the intervals $[k, k+1]$, $k \in \mathbb{Z}$, has support $[0, n+1]$, and is contained in $W^{s,p}$ if and only if $s < n + 1/p$.

Of importance to us is the reproduction of polynomials of degree up to n in the space generated by the translates of B_n: for $m = 0, \ldots, n$, one has

$$x^m := \sum_{k\in\mathbb{Z}} (k^m + Q_{m-1}(k))B_n(x-k), \tag{5.2}$$

where Q_{m-1} are uniquely determined polynomials of degree $m-1$.

Finally, recall that the translates $B_n(\cdot - k)$, $k \in \mathbb{Z}$, constitute a Riesz basis of their span. In particular, there exists a dual spline function $r(x) = \sum r_k B_n(\cdot - k)$ (infinitely supported with exponential decay if $n \geq 1$) such that

$$\langle r(x-k), B_n(\cdot - l)\rangle = \delta_{k,l}. \tag{5.3}$$

Here we want to use the function $\varphi(x) := B_n(x_1)\cdots B_n(x_d)$ (and its shifted and dilated versions $\varphi(h^{-1}\cdot - k)$, $k \in \mathbb{Z}^d$) to generalize the hat function that was used in the example of the previous section. To do so, we also need a proper discretization method that generalizes (4.2) and (4.3). In other word, we want to build an approximation operator

$$P_h f := \sum_{k\in\mathbb{Z}^d} h^{-m}\langle f, \tilde{\varphi}(h^{-1}\cdot - k)\rangle \varphi(h^{-1}\cdot - k) \tag{5.4}$$

onto the spaces

$$V_h := \mathrm{Span}\{\varphi(h^{-1}\cdot - k)\ ;\ k \in \mathbb{Z}^d\}, \tag{5.5}$$

where $\tilde{\varphi}$ is a compactly supported function that was chosen to be $\chi_{[-1/2,1/2]}$ in the example of the previous section. Although it seems natural to choose $\tilde{\varphi} = \chi_{[-1/2,1/2]^d}$,

the construction of $\tilde{\varphi}$ needs to be slightly more refined in order that the operator P_h has some good approximation properties. Such properties are related to the invariance of polynomials of degree up to n under the action of P_h. There are many ways to construct functions $\tilde{\varphi}$ that will yield this property.

In [6] it is shown how to build such a function with the additional prescriptions that P_h is a projector (i.e., $P_h^2 = P_h$) (or equivalently that $\langle \varphi(\cdot - k), \tilde{\varphi}(\cdot - l)\rangle = \delta_{k,l}$), and that the spaces $\widetilde{V}_h$ generated by $\tilde{\varphi}$ satisfy $\widetilde{V}_{2h} \subset \widetilde{V}_h$ (similarly to the spaces V_h).

Here we ignore these additional properties, and we give below a simple criterion for the design of $\tilde{\varphi}$.

LEMMA 5.1. *Let g be any compactly supported univariate function such that $\int g = 1$ and*

$$\int x^m g(x)dx = Q_{m-1}(0) \tag{5.6}$$

for $m = 1, \dots, n$. Then one has the identities

$$\sum_{k\in\mathbb{Z}} \left[\int y^m g(y-k)dy \right] B_n(x-k) = x^m \tag{5.7}$$

for $m = 0, \dots, n$. Defining $\tilde{\varphi}(x) = g(x_1)\cdots g(x_d)$, it follows that the corresponding operator P_h is a projector and satisfies $P_h f = f$ for all f polynomial of degree less than or equal to n in each coordinate.

Proof. We first remark that (5.3) and (5.2) give

$$k^m + Q_{m-1}(k) = \int x^m r(x-k)dx, \quad m = 0, \dots, n. \tag{5.8}$$

Now since

$$\int x^m f(x-k)dx = \sum_{l=0}^{m} \binom{m}{l} (-k)^{m-l} \int x^l f(x)dx, \tag{5.9}$$

the identities $\int x^m g(x)dx = \int x^m r(x)dx$, $m = 0, \dots, n$, imply

$$\int x^m g(x-k)dx = \int x^m g(x-k)dx = k^m + Q_{m-1}(k), \quad m = 0, \dots, n, \tag{5.10}$$

so that (5.7) holds.

Rescaling these identities by a factor h and using the tensor product structure of P_h shows that this operator preserves all polynomials $x_1^{m_1} \cdots x_d^{m_d}$, $d_i \le n$, and thus all polynomials of degree less than or equal to n in each coordinate. ∎

Remark 5.1. According to the previous lemma, the only prescription in the choice of g is its compact support and the value of its $n+1$ first moments. A simple choice consists of choosing g piecewise constant with support $[0, n+1]$, i.e., $g = \sum_{k=0}^{n} g_k B_0(\cdot - k)$: one easily checks that the coefficients g_k are uniquely determined from the moments $\int x^m g(x)$.

We are now ready to describe in full generality the pseudoparticle method. We use the notation $u_t = u(t)$, $u_{t,h} = u_h(t)$... Given the initial value u_0, we thus define the approximation

$$v_{0,h} = P_h u_0 = \sum_{k\in\mathbb{Z}^d} \alpha_k \varphi_h(\cdot - kh), \tag{5.11}$$

where $\varphi_h := h^{-d}\varphi(h^{-1}\cdot)$ and $\alpha_k := \langle u_0, \tilde{\varphi}(h^{-1}\cdot - k)\rangle$. Note that with the choice of g suggested by Remark 5.1, the values of α_k are linear combinations of the integrals of u_0 over cubes of width h in the neighborhood of the grid point kh.

The values of $v_{0,h}$ at the grid points are then given by

$$v_{0,h}(kh) = \sum_{l\in\mathbb{Z}^d} \alpha_l \varphi_h(kh - lh), \tag{5.12}$$

i.e., by a simple convolution of the sequence α_l with a local (and separable) discrete sequence.

As in the example of the previous section, the values $v_{t,h} := v_h(t)$ at the points $x_k(t)$ at time $t > 0$ can be computed, since they are the solutions of the ordinary differential equations (1.4) with initial value $v_{0,h}(kh)$.

We then need a proper interpolation procedure in order to construct the approximation $u_{t,h}$ of the solution at time t from the point values $v_{t,h}(x_k(t))$. This task is more difficult than in the previous example: in the multivariate setting the points of the transported grid are not ordered in a simple way. Yet we can take advantage of the fact that, as h goes to zero, these grid points are locally close to the transport of the initial square grid by the linearized flow.

To do so, we first define

$$S(t) := \sup_{x\in\mathbb{R}^d} \|D\Phi_t(x)\| \ \text{ and } \ \widetilde{S}(t) := \sup_{x\in\mathbb{R}^d} \|(D\Phi_t)^{-1}(x)\|, \tag{5.13}$$

where $\|.\|$ is the standard norm for $d \times d$ matrices. From the assumptions on the coefficients a_i, these quantities stay finite for $t \in [0, T]$.

From now on, we fix $t \in]0, T]$, and we consider the partition of $\mathbb{R}^d$ by the cubes

$$Q^t_{k,h} := (n+1)khS(t) + Q^t_h, \ \ k \in \mathbb{Z}^d, \ \text{ with } \ Q^t_h := [0, (m+1)hS(t)]^d. \tag{5.14}$$

On each of these cubes, we shall define $u_{t,h}$ as a polynomial of degree n in each coordinate.

Note that the choice of the width $(n+1)hS(t)$ ensures that each cube $Q^t_{k,h}$ contains at least $(n+1)^d$ points of the transported grid $\Phi_t(h\mathbb{Z}^d)$ and at most $N(t) := (n+1)^d hS(t)\tilde{S}(t)$ points. The main problem here is that these sets do not generally contain a unisolvent set of $(n+1)^d$ points that allows us to define a unique interpolation polynomial of degree n in each coordinate. However, this problem is solved for h small enough since, as we already mentioned above, the grid points are locally close to a regular mesh.

The polynomial approximation on $Q^t_{k,h}$ will thus be obtained from the values of $v_{t,h}$ at the points $x_k(t)$ that are situated in this cube. Therefore we write

$$u_{t,h} = I^t_{k,h} v_{t,h} \ \text{ on } \ Q^t_{k,h}. \tag{5.15}$$

To define a proper interpolation procedure $I^t_{k,h}$, we first transfer the problem on the unit cube $Q := [0, 1]^d$ through the affine transformation

$$T^t_{k,h} : \ x \mapsto (n+1)hS(t)x + (n+1)khS(t) \tag{5.16}$$

that maps Q onto $Q^t_{k,h}$. The problem is now to find an appropriate construction of a polynomial V of degree n in each variable, given some points $x_1, \ldots, x_q$, $(n+1)^d \leq q \leq N(t)$ in Q and values $y_1, \ldots, y_q$.

We first consider the canonical basis $(g_l)_{l=1,\dots,(n+1)^d}$, $g_l(x) = x_1^{m_1(l)} \cdots x_d^{m_d(l)}$, for the space of multivariate polynomials of degree n in each variable, and we write

$$V(x) = \sum_{l=1,\dots,(n+1)^d} z_l g_l. \tag{5.17}$$

We choose the vector $(z_l)_{l=1,\dots,(n+1)^d}$ to be the image of the vector $(y_1, \dots, y_q)$ by a rectangular matrix $P = (p_{i,j})$ that we construct in the following way:

(i) We solve the problem $\min_P \sum_{i,j} |p_{i,j}|^2$ under the constraints of polynomial exactness: P applied to $(g_l(x_1), \dots, g_l(x_q))$ is equal to the canonical basis vector δ_l corresponding to $V(x) = g_l$. This problem always has a unique solution.

(ii) In the case where there exists a solution P^* such that $\sum_{i,j} |p^*_{i,j}|^2 \le K(t)$, for some $K(t)$ that we specify below, we choose $P = P^*$. Otherwise, we choose $p_{1,1} = 1$ and $p_{i,j} = 0$ otherwise.

Imposing a limitation on $\sum_{i,j} |p^*_{i,j}|^2$ implies a stability property in the L^∞ norm: we have

$$\|V\|_{L^\infty(Q)} \le \widetilde{K}(t) \max_{k=1,\dots,q} |y_k|, \tag{5.18}$$

where $\tilde{K}(t)$ depends only on $K(t)$ and $N(t)$.

The constant $K(t)$ should be chosen large enough to ensure that for h small enough, we are always in the first case of (ii), i.e., $\sum_{i,j} |p^*_{i,j}|^2 \le K(t)$.

This is made possible by the fact that for h small enough, the grid points are locally close to the regular mesh obtained by the transport of the square grid $h\mathbb{Z}^d$ by the linearized flow.

More precisely, if A is a $d \times d$ invertible linear transformation, we consider the unisolvent set of points $(Ax_1, \dots, Ax_{(n+1)^d})$ where

$$\{x_1, \dots, x_{(n+1)^d}\} := \{0, 1/(n+1), 2/(n+1), \dots, n/(n+1)\}^d, \tag{5.19}$$

and we define P_A to be the square matrix that maps the value y_i at these points to the coordinates z_i of the unique Lagrange interpolation polynomial. Then $K(t)$ should be chosen strictly larger than the quantity

$$K_{lim}(t) := \sup_{x \in \mathbb{R}^d} \|P_{D\Phi_t(x)}\|_2^2 \ (< +\infty) \tag{5.20}$$

(for example, take $K(t) := 2K_{lim}(t)$ where $\|P\|_2^2 := \sum_{i,j} |p_{i,j}|^2$). Such a choice ensures that for h small enough, a simple interpolation procedure produces a solution P^* that corresponds to the first case of (ii).

We summarize below the properties that hold for our interpolation procedure.

PROPOSITION 5.2. *The interpolation operator $I^t_{k,h}$ is L^∞-stable, i.e.,*

$$\|u_{t,h}\|_{L^\infty(Q^t_{k,h})} \le C(t) \|v_{t,h}\|_{L^\infty(Q^t_{k,h})}, \tag{5.21}$$

independently of h and k. Moreover, for $h \le h_0(t)$, it is exact for polynomials up to order n in each coordinate.

6. Error estimates. In this section we shall obtain the optimal error estimate for the approximation $u_{t,h}$. For this, we need several preliminary results that concern the approximation by $v_{t,h}$ and should be viewed as the generalization of estimate (4.4)

and (4.8) in section 4. We recall the approximation operator

$$P_h f = \sum_{k\in\mathbb{Z}^d} h^{-d}\langle f, \tilde\varphi(h^{-1}\cdot -k)\rangle \varphi(h^{-1}\cdot -k), \tag{6.1}$$

where φ is the B-spline of order n and $\tilde\varphi$ the corresponding dual function constructed in the previous section. One can easily see that operator is L^p-stable and that its norm in L^p is independent of h: defining p' the conjugate exponent of p (i.e., $1/p+1/p'=1$), we have

$$\begin{aligned}\|P_h f\|_{L^p} &\le Ch^{-d}\|\varphi(h^{-1}\cdot)\|_{L^p}\|(\langle f,\tilde\varphi(h^{-1}\cdot -k)\rangle)_{k\in\mathbb{Z}^d}\|_{\ell^p}\\ &\le Ch^{d(1/p-1)}\|(\langle f,\tilde\varphi(h^{-1}\cdot -k)\rangle)_{k\in\mathbb{Z}^d}\|_{\ell^p}\\ &\le Ch^{d(1/p-1)}\|\tilde\varphi(h^{-1}\cdot)\|_{L^{p'}}\|(\|f\|_{L^p(\mathrm{Supp}(\tilde\varphi(h^{-1}\cdot -k)))})_{k\in\mathbb{Z}^d}\|_{\ell^p}\\ &\le C\|(\|f\|_{L^p(\mathrm{Supp}(\tilde\varphi(h^{-1}\cdot -k)))})_{k\in\mathbb{Z}^d}\|_{\ell^p}\\ &\le C\|f\|_{L^p}.\end{aligned}$$

Here we have used Hölder inequality, and the compact supports of φ and $\tilde\varphi$ which ensure that the support of $\varphi(h^{-1}\cdot -k)$ overlaps a controlled number ($[2n]^d$) of supports of other $\varphi(h^{-1}\cdot -l)$ and similarly for $\tilde\varphi$. We shall reuse this type of argument in the following section.

Here we shall make an important use of fractional Sobolev spaces $W^{s,p}$. Let us recall that when s is not an integer, these spaces can be identified by the Besov spaces $B^s_{p,p}$ for all $p\in[1,+\infty]$ (see, e.g., [16]). In particular, if Q is a cube, the seminorm for $W^{s,p}(Q)$ can be expressed in an equivalent manner by

$$|f|_{W^{s,p}(Q)} = \|(2^{sj}\omega_m(f,2^{-j})_{L^p})_{j\ge 0}\|_{\ell^p}, \tag{6.2}$$

where m is any integer strictly larger than s and $\omega_m(f,t)_{L^p}$ is the mth order L^p modulus of smoothness:

$$\omega_m(f,t)_{L^p} := \sup_{|h|\le t}\|(\Delta_h)^m f\|_{L^p(Q_t)}, \tag{6.3}$$

with $\Delta_h f := f(\cdot)-f(\cdot-h)$ the finite difference operator and $Q_t := \{x\in Q \text{ such that } x, x-h,\dots,x-mh\in Q\}$. Recall also that $\varphi\in W^{s,p}$ if and only if $s<n+1/p$.

LEMMA 6.1. *The approximation operator P_h satisfies the direct estimate*

$$\|u-P_h u\|_{L^p} \le Ch^s|u|_{W^{s,p}} \tag{6.4}$$

for $0<s\le n+1$. Consequently one has, for $t\in[0,T]$,

$$\|u_t - v_{t,h}\|_{L^p} \le C(t)h^s\|u_0\|_{W^{s,p}}. \tag{6.5}$$

Proof. The estimate (6.4) is classical in approximation theory, although it does not always appear in this precise form, so we shall sketch only the proof here.

We define the cubes $J_{k,h} := kh+[0,h]^d$, $k\in\mathbb{Z}^d$. We also define larger cubes $\widetilde{J}_{k,h} := kh+[-nh,(n+1)h]^d$. With the choice of $\tilde\varphi$ supported like φ in $[0,n+1]^d$ as suggested by Remark 5.1, it follows that the value of $P_h f$ on $J_{k,h}$ is only influenced by the value of f on $\widetilde{J}_{k,h}$. In particular, we have a local L^p-stability estimate

$$\|P_h f\|_{L^p(J_{k,h})} \le C\|f\|_{L^p(\widetilde{J}_{k,h})}. \tag{6.6}$$

From the Deny–Lions theorem (see, e.g., [4]) and a classical scaling argument, we have the estimate

$$\inf_{g\in\Pi_n} \|f-g\|_{L^p(\widetilde{J}_{k,h})} \le Ch^s |f|_{W^{s,p}(\widetilde{J}_{k,h})} \tag{6.7}$$

for the local approximation by polynomials of global degree n. For $k \in \mathbb{Z}^d$, we consider $g_k \in \Pi_n$ such that $\|f-g_k\|_{L^p(\widetilde{J}_{k,h})} \le 2\inf_{g\in\Pi_n} \|f-g\|_{L^p(\widetilde{J}_{k,h})}$. Combining the polynomial reproduction property of P_h and the local L^p-stability property (6.6) together with the estimate (6.7), we obtain

$$\begin{aligned} \|f-P_hf\|_{L^p(J_{k,h})} &\le \|f-g_k\|_{L^p(J_{k,h})} + \|P_hf-P_hg_k\|_{L^p(J_{k,h})} \\ &\le C\|f-g_k\|_{L^p(\widetilde{J}_{k,h})} \le Ch^s|f|_{W^{s,p}(\widetilde{J}_{k,h})}. \end{aligned}$$

Elevating to the power p and summing on k (or taking the supremum in the case $p=\infty$) yields (6.4), using the fact that a cube $\widetilde{J}_{k,h}$ overlaps a controlled number ($[4n]^d$) of other cubes $\widetilde{J}_{l,h}$.

It follows that

$$\|u_0 - v_{0,h}\|_{L^p} \le Ch^s \|u_0\|_{W^{s,p}}, \tag{6.8}$$

which yields (6.5) by the stability of the linear evolution operator. ∎

For the following results, we assume that $h \le 1$ and we denote by j_h the positive integer such that $2^{j_h}h \in]1/2, 1]$.

LEMMA 6.2. *The functions in the space* $V_h = \mathrm{Span}\{\varphi_h(\cdot - kh)\ ;\ k\in\mathbb{Z}^d\}$ *satisfy the inverse estimate*

$$\|f_h\|_{W^{s',p}} \le Ch^{s-s'}\|f_h\|_{W^{s,p}}, \quad f_h \in V_h, \tag{6.9}$$

for $0 \le s \le s' < n + 1/p$. *In particular we have*

$$\|v_{0,h}\|_{W^{s',p}} \le Ch^{s-s'}\|v_{0,h}\|_{W^{s,p}}. \tag{6.10}$$

Proof. We first prove this result when $s = 0$, i.e., with the L^p norm on the left side of (6.9). We first consider the case where s' is an integer. Here it suffices to remark that if $m = (m_1, \dots, m_d)$ is such that $m_1 + \cdots + m_d = s'$, and if $f_h = \sum_{k\in\mathbb{Z}^d} c_k\varphi(h^{-1}\cdot - k) \in V_h$, we have

$$\begin{aligned} \|\partial^m f_h\|_{L^p} &\le \|h^{-s'} \sum_{k\in\mathbb{Z}^d} c_k[\partial^m\varphi](h^{-1}\cdot - k)\|_{L^p} \\ &\le Ch^{-s'}h^{d/p}\|(c_k)_{k\in\mathbb{Z}^d}\|_{\ell^p}, \end{aligned}$$

where the last inequality makes use of the fact that the support of $\varphi(h^{-1}\cdot - k)$ overlaps a controlled number ($[2n]^d$) of supports of other $\varphi(h^{-1}\cdot - l)$.

In order to obtain (6.9) we use the reverse L^p-stability property of the B-splines which yields

$$h^{d/p}\|(c_k)_{k\in\mathbb{Z}^d}\|_{\ell^p} \le C\|f_h\|_{L^p}. \tag{6.11}$$

One possible technique to prove (6.11) is to use the the compactly supported dual function $\tilde{\varphi}^c$ (such as those built in [6]) that satisfies $\langle\varphi(\cdot - k), \tilde{\varphi}(\cdot - l)\rangle = \delta_{k,l}$ and to evaluate the coefficients c_k using Hölder's inequality as follows:

$$\begin{aligned} |c_k| &= |h^{-d}\langle f_h, \tilde{\varphi}^c(h^{-1}\cdot - k)\rangle| \\ &\le C\|f_h\|_{L^p(\mathrm{Supp}(\tilde{\varphi}^c(h^{-1}\cdot - k)))}. \end{aligned}$$

This yields (6.11) by elevating to the power p and summing on k (or taking the supremum in the case $p = \infty$).

The case where s is not an integer can be treated by an interpolation argument, except for the values of s between n and $n + 1/p$. For such values, we need a specific argument that we give here.

For this, we need to evaluate $\omega_m(f_h, 2^{-j})_{L^p}$. We first remark that

$$\omega_m(f_h, 2^{-j})_{L^p} \leq C\|(c_k)_{k\in\mathbb{Z}^d}\|_{\ell^p} h^d \omega_m(\varphi_h, 2^{-j})_{L^p}, \tag{6.12}$$

so that

$$|f_h|_{W^{s',p}} \leq C\|(c_k)_{k\in\mathbb{Z}^d}\|_{\ell^p} h^d |\varphi_h|_{W^{s',p}}. \tag{6.13}$$

We thus want to evaluate $|\varphi_h|_{W^{s',p}}$. For $j \leq j_h$, we use the crude estimate

$$\omega_m(\varphi_h, 2^{-j})_{L^p} \leq \|\varphi_h\|_{L^p} \leq Ch^{d(1/p-1)}. \tag{6.14}$$

For $j \geq j_h$, we exploit the fact that $\varphi \in W^{s',p}$ to obtain the estimate

$$\omega_m(\varphi_h, 2^{-j})_{L^p} \leq C2^{s'(j_h-j)} h^{d(1/p-1)} \varepsilon_{j-j_h}, \tag{6.15}$$

where $(\varepsilon_n)_{n\geq 0}$ is an ℓ^p sequence. It follows that

$$\begin{aligned} |\varphi_h|_{W^{s',p}} &\leq \|(2^{js'}\omega_m(f_h, 2^{-j})_{L^p})_{j\geq 0}\|_{\ell^p} \\ &\leq \|(2^{js'}\omega_m(f_h, 2^{-j})_{L^p})_{0\leq j\leq j_h}\|_{\ell^p} + \|(2^{js'}\omega_m(f_h, 2^{-j})_{L^p})_{j\geq j_h}\|_{\ell^p} \\ &\leq C2^{j_h s'} h^{d(1/p-1)} \leq Ch^{-s'+d(1/p-1)}. \end{aligned}$$

Combining with (6.13), we obtain

$$|f_h|_{W^{s',p}} \leq C\|(c_k)_{k\in\mathbb{Z}^d}\|_{\ell^p} h^{-s'+d/p}, \tag{6.16}$$

which together with (6.11) yields the inverse estimate.

Finally, in order to treat the case $0 < s < s'$, we use a multiscale decomposition of f_h to obtain

$$\begin{aligned} \|f_h\|_{W^{s',p}} &\leq \|f_h - P_{2h}f_h\|_{W^{s',p}} + \|P_{2h}f_h - P_{4h}f_h\|_{W^{s',p}} + \cdots \\ &\quad + \|P_{2^{j_h-1}h}f_h - P_{2^{j_h}h}f_h\|_{W^{s',p}} + \|P_{2^{j_h}h}f_h\|_{W^{s',p}} \\ &\leq C[h^{-s'}\|f_h - P_{2h}f_h\|_{L^p} + (2h)^{-s'}\|P_{2h}f_h - P_{4h}f_h\|_{L^p} + \cdots \\ &\quad + \|P_{2^{j_h}h}f_h\|_{L^p}] \\ &\leq C\left[\|f_h\|_{L^p} + \sum_{j=1}^{j_j} (2^j h)^{-s'}\|f_h - P_{2^j h}f_h\|_{L^p}\right] \\ &\leq C\left[\|f_h\|_{L^p} + \left(\sum_{j=1}^{j_j} (2^j h)^{-s'+s}\right)|f_h|_{W^{s,p}}\right] \\ &\leq Ch^{s-s'}\|f_h\|_{W^{s,p}}, \end{aligned}$$

where we have successively used the inverse estimate for $s = 0$, the L^p-stability of the projectors P_h, and the direct estimate of Lemma 6.1. □

LEMMA 6.3. *The approximation operator is stable in $W^{s,p}$, i.e.,*

$$\|P_h f\|_{W^{s,p}} \leq C\|f\|_{W^{s,p}}, \tag{6.17}$$

for $0 < s < n + 1/p$.

Proof. In the case where s is an integer we can use a particular property of the B-spline: if $s \leq n$, one has

$$B_n^{(s)} = (\Delta_1)^s B_{n-s}, \tag{6.18}$$

where $\Delta_1 f = f(\cdot) - f(\cdot - 1)$ (this is easily proved by induction on s, using the definition of B-splines). If we now consider the operator

$$A_0 f := \sum_{k \in \mathbb{Z}} \langle f, r(\cdot - k) \rangle B_n(\cdot - k), \tag{6.19}$$

acting on univariate function, we can derive from (6.18) and integration by part the following commutation formula:

$$(A_0 f)^{(s)} = A_s f^{(s)}, \text{ where } A_s f := \sum_{k \in \mathbb{Z}} \langle f, r_s(\cdot - k) \rangle B_{n-s}(\cdot - k), \tag{6.20}$$

by r_s is obtained from $r = r_0$ by

$$r_{m+1}(x) := -\int_x^{x+1} r_m(t) dt. \tag{6.21}$$

In particular, the functions r_m are compactly supported. After tensor product and rescaling of a factor h, we obtain for $m = (m_1, \ldots, m_d)$ such that $m_1 + \cdots + m_d = s$,

$$\partial^m (P_h f) = \sum_{k \in \mathbb{Z}^d} h^{-d} \langle \partial^m f, \tilde{\varphi}_m(h^{-1} \cdot - k) \rangle \varphi_m(h^{-1} \cdot - k), \tag{6.22}$$

where $\varphi_m(x) = B_{n-m_1}(x_1) \cdots B_{n-m_d}(x_d)$ and $\tilde{\varphi}(x) = r_{m_1}(x_1) \cdots r_{m_d}(x_d)$. Using the same arguments as for the L^p-stability of P_h we thus obtain

$$\|\partial^m (P_h f)\|_{L^p} \leq C \|\partial^m f\|_{L^p}, \tag{6.23}$$

independently of h.

As for the inverse estimate of the previous lemma, the case where s is not an integer can be treated by an interpolation argument, except for the values of s between n and $n + 1/p$. We thus give a specific argument here, although a bit involved.

When s is not an integer, the spaces $W^{s,p} := B^s_{p,p}$ also have a simple characterization through approximation properties by stable L^p-projectors onto the spaces V_h.

Here we shall use the specific projector P_h^c, associated with the dual function φ^c constructed in [6], such that the spaces $\widetilde{V}_h$ generated by $\tilde{\varphi}$ satisfy $\widetilde{V}_{2h} \subset \widetilde{V}_h$. This last property implies that

$$P_2^c h P_h^c f = P_2 h f. \tag{6.24}$$

For such projectors and $a \in [1/2, 1[$, we have the norm equivalence

$$\|f\|_{B^s_{p,p}} \sim \|P_a^c\|_{L^p} + \|(2^{sj} \|P^c_{2^{j+1}a} f - P^c_{2^j a} f\|_{L^p})_{j \geq 0}\|_{\ell^p}, \tag{6.25}$$

with constants that do not depend on a. We refer to [5], [8], and [9] for a description of the general mechanism (involving direct estimates, inverse estimates, and interpolation of function spaces) that yields such norm equivalences.

Combining (6.24) with (6.25), we obtain

$$\begin{aligned}\|P_h^c f\|_{B^s_{p,p}} &\leq C\|P^c_{2^{j_h}h}\|_{L^p} + \|(2^{sj}\|P^c_{2^{j+1}h}f - P^c_{2^j h}f\|_{L^p})_{0\leq j\leq j_h}\|_{\ell^p}\\ &\leq C\|f\|_{B^s_{p,p}}.\end{aligned}$$

We have thus proved the $W^{s,p}$-stability for P_h^c, uniformly in h. In order to see that the same property holds for P_h, we write

$$\begin{aligned}\|P_h f\|_{B^s_{p,p}} &\leq \|P_h^c f\|_{B^s_{p,p}} + \|P_h^c f - P_h f\|_{B^s_{p,p}}\\ &\leq C[\|f\|_{B^s_{p,p}} + h^{-s}\|P_h^c f - P_h f\|_{L^p}]\\ &\leq C[\|f\|_{B^s_{p,p}} + h^{-s}(\|P_h^c f - f\|_{L^p} + \|P_h f - f\|_{L^p})]\\ &\leq C\|f\|_{B^s_{p,p}},\end{aligned}$$

where we have used the inverse estimate of Lemma 6.2 together with the direct estimate for both P_h and P_h^c. □

If we combine Lemma 6.2 and Lemma 6.3 together with the stability of the evolution operator, we obtain an inverse estimate of the type

$$\|v_{t,h}\|_{W^{s',p}} \leq C(t)h^{s-s'}\|u_0\|_{W^{s,p}} \tag{6.26}$$

for $t \in [0,T]$ and $0 \leq s \leq s' < n - 1 + 1/p$. We are now ready to prove the main result. At this point we make a technical assumption on the order of the B-spline: n is large enough so that $d/p \leq n + 1/p$.

THEOREM 6.4. *The approximation of u_t by the pseudoparticle method satisfies*

$$\|u_t - u_{t,h}\|_{L^p} \leq C(t)h^s\|u_0\|_{W^{s,p}} \tag{6.27}$$

for $0 < s < n - 1 + 1/p$.

Proof. On each cube $Q^t_{k,h}$, we consider a polynomial $p^t_{k,h} \in \Pi_n$ such that

$$\|v_{t,h} - p^t_{k,h}\|_{L^p(Q^t_{k,h})} \leq 2\inf_{g\in\Pi_n}\|v_{t,h} - g\|_{L^p(Q^t_{k,h})} \leq Ch^{s'}|v_{t,h}|_{W^{s',p}(Q^t_{k,h})} \tag{6.28}$$

and we set $r^t_{k,h} = v_{t,h} - p^t_{k,h}$. According to Lemma 5.3, we have

$$v_{t,h} - u_{t,h} = r^t_{k,h} - I^t_{k,h}r^t_{k,h} \tag{6.29}$$

for $h \leq h_0(t)$. We thus have

$$\begin{aligned}\|v_{t,h} - u_{t,h}\|_{L^p(Q^t_{k,h})} &\leq C(t)h^{d/p}\|v_{t,h} - u_{t,h}\|_{L^\infty(Q^t_{k,h})}\\ &= C(t)h^{d/p}\|r^t_{k,h} - I^t_{k,h}r^t_{k,h}\|_{L^\infty(Q^t_{k,h})}\\ &\leq C(t)h^{d/p+\sigma}|r^t_{k,h}|_{W^{\sigma,\infty}(Q^t_{k,h})}.\end{aligned}$$

Here the direct estimate for $I^t_{k,h}$ uses the L^∞-stability of this operator (Proposition 5.2) together with the Deny–Lions theorem, by the same argument as in Lemma 6.1.

Now choose s' such that $\max\{d/p, s\} < s' < n+1/p$, and such that $\sigma := s' - d/p > 0$ is not an integer. The Sobolev embedding of $W^{s',p}$ into $W^{\sigma,\infty}$ gives after rescaling

$$|r^t_{k,h}|_{W^{\sigma,\infty}(Q^t_{k,h})} \leq C(t)\Big(|r^t_{k,h}|_{W^{s',p}(Q^t_{k,h})} + h^{-s'}\|r^t_{k,h}\|_{L^p(Q^t_{k,h})}\Big). \tag{6.30}$$

Using that $|p^t_{k,h}|_{W^{s',p}(Q^t_{k,h})} = 0$ and (6.28), we thus obtain

$$|r^t_{k,h}|_{W^{\sigma,\infty}(Q^t_{k,h})} \leq C(t)|v_{t,h}|_{W^{s',p}(Q^t_{k,h})}. \tag{6.31}$$

Combining (6.31) with the estimate for $\|v_{t,h} - u_{t,h}\|_{L^p(Q^t_{k,h})}$, we thus obtain

$$\|v_{t,h} - u_{t,h}\|_{L^p(Q^t_{k,h})} \leq C(t)h^{s'}|v_{t,h}|_{W^{s',p}(Q^t_{k,h})}. \tag{6.32}$$

Together with (6.26), this yields

$$\|v_{t,h} - u_{t,h}\|_{L^p(Q^t_{k,h})} \leq C(t)h^s\|u_0\|_{W^{s,p}(Q^t_{k,h})}, \tag{6.33}$$

and thus

$$\|v_{t,h} - u_{t,h}\|_{L^p} \leq C(t)h^s\|u_0\|_{W^{s,p}}, \tag{6.34}$$

by elevating to the power p and summing on k (or taking the supremum in the case $p = \infty$). Combining (6.34) and (6.5), we finally obtain the optimal estimate (6.27). □

Remark 6.1. The pseudoparticle method is by essence not conservative: if $a_0 = f = 0$, in which case $\int u_t(x)dx = \int u_0(x)dx$, we do not have $\int u_{t,h}(x)dx = \int u_0(x)dx$ in contrast to the classical particle method. However, we can always apply a global correction to the solution $u_{t,h}$ in such a way that this conservation holds. For example, we can add a correction of the form $A\Phi(x)$, where Φ is smooth, globally supported, and such that $\int \Phi(x)dx = 1$ and where $A = \int(u_0(x) - u_{t,h}(x))dx$. We then remark that the estimate (6.27) of Theorem 6.4 yields $|A| \leq C(t)h^s\|u_0\|_{W^{s,p}}$, so that the corrected solution will also satisfy an optimal estimate of the same type.

Remark 6.2. The approximation of u_t by $u_{t,h}$ is optimal in L^p norms. However, $u_{t,h}$ is discontinuous. We end by showing that by applying the operator P_h, one can always regularize this approximation in order to obtain optimal estimates in smoother norms.

THEOREM 6.5. *The regularized approximation of* u_t *by* $\tilde{u}_{t,h} := P_h u_{t,h}$ *satisfies*

$$\|u_t - \tilde{u}_{t,h}\|_{W^{s',p}} \leq C(t)h^{s-s'}\|u_0\|_{W^{s,p}} \tag{6.35}$$

for $0 < s' < s < n + 1/p$.

Proof. We write

$$\|u_t - \tilde{u}_{t,h}\|_{W^{s',p}} \leq \|P_h u_t - P_h u_{t,h}\|_{W^{s',p}} + \|u_t - P_h u_t\|_{W^{s',p}}. \tag{6.36}$$

The first term satisfies

$$\begin{aligned}\|P_h u_t - P_h u_{t,h}\|_{W^{s',p}} &\leq Ch^{-s'}\|P_h u_t - P_h u_{t,h}\|_{L^p} \\ &\leq Ch^{-s'}\|u_t - u_{t,h}\|_{L^p} \\ &\leq C(t)h^{s-s'}\|u_0\|_{W^{s,p}}\end{aligned}$$

by the inverse estimate of Lemma 6.2, the L^p-stability of P_h, and the L^p error estimate of Theorem 6.4. For the second term, we need a direct estimate in $W^{s',p}$ for P_h. This is easily obtained by the following multiscale technique:

$$\begin{aligned}\|f - P_h f\|_{W^{s',p}} &\leq \sum_{j\geq 0}\|P_{2^{-j}h}f - P_{2^{-(j+1)}h}f\|_{W^{s',p}} \\ &\leq C\sum_{j\geq 0}(2^{-j}h)^{-s'}\|P_{2^{-j}h}f - P_{2^{-(j+1)}h}f\|_{L^p} \\ &\leq C\sum_{j\geq 0}(2^{-j}h)^{-s'}\|P_{2^{-j}h}f - f\|_{L^p} \\ &\leq C\sum_{j\geq 0}(2^{-j}h)^{s-s'}|f|_{W^{s,p}} \\ &\leq Ch^{s-s'}|f|_{W^{s,p}},\end{aligned}$$

where we have used both the direct and inverse estimates of Lemma 6.1 and Lemma 6.2. We thus have the optimal estimate for both terms in (6.36), which concludes the proof. □

REFERENCES

[1] J.T. Beale, *On the accuracy of vortex methods at large time*, Computational Fluid Dynamics and Reacting Gas Flows, B. Engquist, A. Majda, and M. Luskin, eds., Springer-Verlag, New York, Berlin, 1988.

[2] C. de Boor and G. Fix, *Approximation from shift-invariant subspaces* of $L_2(\mathbb{R}^d)$, Trans. Amer. Math. Soc., 341 (1973), pp. 787–806.

[3] P. Choquin, Ph.D. thesis, Dept. de Mathématiques Appliquées, Ecole Polytechnique, 91128 Palaiseau, France, 1993.

[4] P.G. Ciarlet, *Basic error estimate in the finite element methods*, Handbook of Numerical Analysis, Vol. 2, P.G. Ciarlet and J.-L. Lions, eds., Elsevier, Amsterdam, 1985.

[5] A. Cohen, *Wavelet methods in numerical analysis*, Handbook of Numerical Analyis, Vol. 7, P.G. Ciarlet and J.-L. Lions, eds., Elsevier, Amsterdam, 2000.

[6] A. Cohen, I. Daubechies, and J.-C. Feauveau, *Biorthogonal bases of compactly supported wavelets*, Comm. Pure Appl. Math. 45 (1992), pp. 485–560.

[7] G.H. Cottet and P. Koumoutsakos, *Vortex Method: Theory and Practice*, Cambridge University Press, Cambridge, UK, 1999.

[8] W. Dahmen *Stability of multiscale transformations*, J. Fourier Anal. Appl., 2 (1996), pp. 341–361.

[9] R. DeVore and G. Lorentz, *Constructive Approximation*, Springer-Verlag, Berlin, 1993.

[10] R.T. Glassey, *The Cauchy Problem in Kinetic Theory*, SIAM, Philadelphia, 1996.

[11] E. Gustafson and J. Sethian, eds., *Vortex Methods and Vortex Motion*, SIAM, Philadelphia, 1991.

[12] T. Y. Hou, *Convergence of a Variable Blob Vortex Method for the Euler and Navier-Stokes Equations*, SIAM J. Numer. Anal., 27 (1990), pp. 1387–1404.

[13] M. Junk, S. Mas Gallic, and P.-A. Raviart, *Introduction to Vortex Methods*, SMAI series, Springer-Verlag, Berlin, to appear.

[14] P.-A. Raviart, *An analysis of particle methods*, in Numerical Methods in Fluid Dynamics, Lecture Notes in Math. 1127, Springer-Verlag, Berlin, 1983, pp. 243–324.

[15] G. Strang and G. Fix, *Fourier analysis of the finite element method in Ritz-Galerkin theory*, Stud. Appl. Math. 48 (1969), pp. 265–273.

[16] H. Triebel, *Theory of Function Spaces*, Birkhauser, Basel, 1983.

SIAM J. MATH. ANAL.
Vol. 32, No. 3, pp. 637–650

STABILITY WITH RESPECT TO PSEUDODIFFERENTIAL PERTURBATIONS OF SOME NONLINEAR DIFFUSIVE EQUATIONS*

MICHELLE SCHATZMAN†

Abstract. Let $g^\varepsilon(k)$ be an even function of $k \in \mathbb{Z}$ which satisfies the inequality

$$\forall k \in \mathbb{Z}, \quad g^\varepsilon(k) \le \rho - \nu|k|^2.$$

Assume, moreover, that $\forall k$, $g^\varepsilon(k)$ tend to a limit $g^0(k)$ as ε tends to 0. Define a linear operator L^ε on periodic functions of period 2π by

$$\big(L^\varepsilon u\big)\widehat{\ }(k) = g^\varepsilon(k)\widehat{u}(k).$$

Joulin has asked whether the solution of

$$u_t^\varepsilon = L^\varepsilon u^\varepsilon - \big(u_x^\varepsilon\big)^2/2$$

converges to the solution of the analogous problem for $\varepsilon = 0$. It is proved here that the answer is positive. Such a positive answer is a means of validating a number of theoretical procedures in the analysis of nonlinear phenomena and particularly of combustion phenomena.

Key words. pseudodifferential operators, stability, semilinear evolution equations, combustion models

AMS subject classifications. 35S10, 35B25, 80A25

PII. S0036141099357860

1. Presentation of the problem.

Denote by $\mathbb{T} = \mathbb{R}/2\pi\mathbb{Z}$ the torus of length 2π and by $\mathcal{D}'_\sharp$ the set of distributions over $\mathbb{T}$, i.e., the set of periodic distributions of period 2π. The space L^p over $\mathbb{T}$ will be denoted by $L^p_\sharp$, and the Sobolev space of functions with s derivatives in $L^2_\sharp$ will be likewise denoted by $H^s_\sharp$. The Fourier coefficient $\widehat{u}(k)$ is defined for any distribution u on $\mathbb{T}$ and any integer $k \in \mathbb{Z}$ by

$$\widehat{u}(k) = \langle u, \exp(-ik\cdot)\rangle.$$

Let ν be a strictly positive number and let ρ be a real number; let $\mathcal{G}$ be the class of even functions from $\mathbb{Z}$ to $\mathbb{R}$ which satisfy the estimate

$$\forall k \in \mathbb{Z}, \quad g(k) \le \rho - \nu|k|^2. \tag{1.1}$$

To each function $g \in \mathcal{G}$, we associate a pseudodifferential linear operator L by

$$\big(Lu\big)\widehat{\ }(k) = g(k)\widehat{u}(k). \tag{1.2}$$

Let g^ε be a sequence of functions belonging to $\mathcal{G}$, which converge point-wise to a limit g^0 in the following sense:

$$\forall k \in \mathbb{Z}, \quad \lim_{\varepsilon\to 0} g^\varepsilon(k) = g^0(k). \tag{1.3}$$

*Received by the editors June 18, 1999; accepted for publication (in revised form) August 22, 2000; published electronically October 20, 2000.
http://www.siam.org/journals/sima/32-3/35786.html
†MAPLY, UMR 5585 CNRS Université Claude Bernard – Lyon 1, 69622 Villeurbanne Cedex, France (schatz@maply.univ-lyon1.fr).

No further assumption is made on the convergence.

The pseudodifferential operator associated with g^ε is denoted by L^ε.

Joulin has asked the following question: is it true that the solution u^ε of the initial value problem

$$\begin{aligned} &\frac{\partial u^\varepsilon}{\partial t} = L^\varepsilon u^\varepsilon - \frac{1}{2}\left(\frac{\partial u^\varepsilon}{\partial x}\right)^2, \quad (x,t) \in \mathbb{T} \times (0,T), \\ &u^\varepsilon(x,0) = \phi(x), \quad x \in \mathbb{T} \end{aligned} \tag{1.4}$$

converges to the solution of the limiting initial value problem

$$\begin{aligned} &\frac{\partial u^0}{\partial t} = L^0 u^0 - \frac{1}{2}\left(\frac{\partial u^0}{\partial x}\right)^2, \quad (x,t) \in \mathbb{T} \times (0,T), \\ &u^0(x,0) = \phi(x), \quad x \in \mathbb{T}, \end{aligned} \tag{1.5}$$

as ε tends to 0?

The motivation for this question comes from a rather general procedure for analyzing nonlinear phenomena in physics and specifically combustion phenomena [3], [4], [1]. In these analyses, one assumes that the behavior of the linear part of the problem is described in Fourier variable: an asymptotic is assumed on the high modes behavior of this linear operator.

Practically, a specific form of the linear operator is assumed in order to simplify the mathematical procedure. However, it would be important to found this procedure more rigorously; after all, if it were unfounded, it would mean that the results would be strongly dependent on the details of the behavior of the linear operator.

The perturbation procedure consists of expanding the symbol g^ε of the operator under study, usually for small wave numbers, assuming that the symbol satisfies reasonable assumptions for large wave numbers, so that the model will be well posed; in particular, it would be very awkward to make an assumption of the form

$$|k| - \nu|k|^2 - \varepsilon|k|^3 \le g^\varepsilon(k) \le |k| - \nu|k|^2 + \varepsilon|k|^3, \tag{1.6}$$

since the rightmost inequality does not guarantee that the symbol g^ε is bounded for large wave numbers.

On the other hand, we could replace the respective coefficients $-\varepsilon$ and ε of the cubic terms in (1.6), respectively, by $-\mu - \varepsilon$ and $-\mu + \varepsilon$, with μ a positive number, and the theory presented here would apply in a straightforward manner.

The diffusion assumption means that estimate (1.1) holds, i.e., $-g^\varepsilon(k)$ has at least quadratic growth at infinity. It is mathematically possible to consider more general behaviors for the symbol, but it may be not extremely useful from the point of view of the applications.

The purpose of this paper is to show that the answer to Joulin's question is positive: if the initial data ϕ belong to $H^1_\sharp$, then u^ε tends to u^0 in $C^0\big([0,T];H^1_\sharp\big)$, for all finite T, as ε tends to 0.

The outline of the proof is as follows: the first step consists of obtaining an existence proof for (1.4), with bounds in $C^0\big([0,T];H^1(\mathbb{T})\big)$ which are independent of $\varepsilon \in [0,1]$. Once these estimates are obtained, we use an almost straightforward Gronwall lemma argument to conclude: the almost refers to the fact that we have to use an integral equation whose kernel contains a factor $t^{-3/4}$, so that a bit of care is required to attain the desired conclusion.

The difficulty comes from the fact that the only obvious estimates that we can obtain on $\exp(tL^\varepsilon)$ are in $L^2_\sharp$ and $H^s_\sharp$; since the nonlinear term has its values in $L^1_\sharp$, we decompose $\exp\bigl(tL^\varepsilon\bigr)$ as follows:

$$\exp\bigl(tL^\varepsilon\bigr) = \exp\bigl(tM^\varepsilon\bigr)\exp\bigl(t\nu\partial_{xx}/2\bigr);$$

the operator M^ε is defined as

$$M^\varepsilon = L^\varepsilon - \nu\partial_{xx}/2$$

and it has essentially the same properties as L^ε; but the operator ∂_{xx} is very well understood, and, in particular, an elementary calculation gives the norm of $\exp\bigl(t\nu\partial_{xx}/2\bigr)$ and of $\partial_x \exp\bigl(t\nu\partial_{xx}/2\bigr)$ as operators from $L^1_\sharp$ to $L^2_\sharp$. This calculation accounts for the fact that the kernel contains a term in $t^{-3/4}$.

It should be noted that the technique presented here is strongly dependent on the fact that the spatial dimension is equal to 1 and the growth of the nonlinearity is at most quadratic.

2. Existence and estimates. In this section, we prove an existence theorem with uniform bounds for a linear operator L defined by (1.2) and a symbol $g \in \mathcal{G}$.

THEOREM 2.1. *Assume that ϕ belongs to $H^2_\sharp$. Then, $\forall T > 0$, there exists a unique solution u of* (1.4)*, which has the following functional property:*

$$u \in C^0\bigl([0,T], H^2_\sharp\bigr) \cap C^1\bigl([0,T]; L^2_\sharp\bigr);$$

Moreover, the following estimate holds:

$$\sup_{g\in\mathcal{G}} \max_{0\le t\le T} |u(\cdot,t)|_{H^2_\sharp} < +\infty.$$

We first prove some auxiliary results on the semigroup $\exp\bigl(tL\bigr)$ and on an integral equation involving it.

LEMMA 2.2. *The operator L generates a holomorphic semigroup in $H^s_\sharp$. This semigroup is denoted by $\exp\bigl(tL\bigr)\forall s \ge 0$, and it satisfies the estimates*

$$\bigl\|\exp\bigl(tL\bigr)\bigr\|_{\mathcal{L}(H^s_\sharp)} \le e^{\rho t}. \tag{2.1}$$

Moreover, there exists a constant C such that $\forall t \ge 0$

$$\bigl\|\exp\bigl(tL\bigr)\bigr\|_{\mathcal{L}(L^1_\sharp, L^2_\sharp)} \le \frac{Ce^{\rho t}}{t^{1/4}}, \tag{2.2}$$

$$\bigl\|\partial_x \exp\bigl(tL\bigr)\bigr\|_{\mathcal{L}(L^1_\sharp, L^2_\sharp)} \le \frac{Ce^{\rho t}}{t^{3/4}}. \tag{2.3}$$

Proof. As is obvious in its Fourier representation, L is self-adjoint and bounded from above in the spaces $H^s_\sharp$; thus it generates a holomorphic semigroup $\exp\bigl(tL\bigr)$ in each of these spaces. The estimate (2.1) is clear in Fourier representation.

Denote the heat kernel by

$$\mathcal{E}(x,t) = \frac{1}{\sqrt{4\pi t}} \exp\left(-\frac{x^2}{4t}\right).$$

A straightforward calculation gives

$$|\mathcal{E}(\cdot,t)|_{L^2(\mathbb{R})} \le Ct^{-1/4},$$
$$|\partial_x \mathcal{E}(\cdot,t)|_{L^2(\mathbb{R})} \le Ct^{-3/4}.$$

Therefore, if z belongs to $L^1_\sharp$, the following estimates hold:

$$|\mathcal{E}(\cdot,t) * z|_{L^2_\sharp} \le Ct^{-1/4} |z|_{L^1_\sharp}, \tag{2.4}$$

$$|\partial_x \mathcal{E}(\cdot,t) * z|_{L^2_\sharp} \le Ct^{-3/4} |z|_{L^1_\sharp}. \tag{2.5}$$

We decompose L as

$$L = \nu\partial_{xx}/2 + M,$$

where M is given in Fourier space by

$$\big(Mu\big)\widehat{\ } = \big(g(k) + \nu k^2/2\big)\widehat{u}(k). \tag{2.6}$$

Since M and ∂_x commute, it is immediate that

$$\big\|\exp\big(tL\big)\big\|_{\mathcal{L}(L^1_\sharp, L^2_\sharp)} \le \big\|\exp\big(tM\big)\big\|_{\mathcal{L}(L^2_\sharp)} \big\|\exp\big(\nu t\partial_{xx}\big)\big\|_{\mathcal{L}(L^2_\sharp, L^1_\sharp)}.$$

By analogy with (2.1), we have

$$\big\|\exp\big(tM\big)\big\|_{\mathcal{L}(L^2_\sharp)} \le e^{\rho t}.$$

When we combine the above relation with (2.4) we find (2.2); we obtain (2.3) analogously. □

We define now for all functions u and v in $C^0\big([0,T]; L^2_\sharp\big)$ two bilinear forms

$$B_0(u,v)(\cdot,t) = \int_0^t \exp\big((t-s)L\big)\big[u(\cdot,s)v(\cdot,s)\big]\, ds, \tag{2.7}$$

$$B_1(u,v)(\cdot,t) = \int_0^t \partial_x \exp\big((t-s)L\big)\big[u(\cdot,s)v(\cdot,s)\big]\, ds. \tag{2.8}$$

The first properties of these bilinear forms are described in the following lemma.

LEMMA 2.3. *The bilinear forms B_0 and B_1 map $C^0\big([0,T], L^2_\sharp\big)^2$ to $C^0\big([0,T], L^2_\sharp\big)$, and there exists a constant C such that the following estimate holds:*

$$|B_0(u,v)(\cdot,t)|_{L^2_\sharp} \le C \int_0^t e^{\rho(t-s)}(t-s)^{-1/4} |u(\cdot,s)|_{L^2_\sharp} |v(\cdot,s)|_{L^2_\sharp}\, ds, \tag{2.9}$$

$$|B_1(u,v)(\cdot,t)|_{L^2_\sharp} \le C \int_0^t e^{\rho(t-s)}(t-s)^{-3/4} |u(\cdot,s)|_{L^2_\sharp} |v(\cdot,s)|_{L^2_\sharp}\, ds. \tag{2.10}$$

Proof. Estimate (2.9) is an immediate consequence of (2.2), and estimate (2.10) is an immediate consequence of (2.3). There remains to prove that the mappings $t \mapsto B_i(u,v)(\cdot,t)$ are continuous for $i = 0, 1$. Consider, for instance, the case $i = 1$, the case $i = 0$ being analogous. We observe that for $h > 0$,

$$\begin{aligned} &B_1(u,v)(\cdot,t+h) - B_1(u,v)(\cdot,t) \\ &= \int_0^t \partial_x \exp(sL)\big[(uv)(\cdot,t+h-s) - (uv)(\cdot,t-s)\big]\, ds \\ &\quad + \int_t^{t+h} \partial_x \exp(sL)\big[(uv)(\cdot,t+h-s)\big]\, ds. \end{aligned} \tag{2.11}$$

By uniform continuity of u and v from $[0,T]$ to $H^1_\sharp$, we can see that

$$\max_{0\le t\le T-h} |u(\cdot,t+h-s)v(\cdot,t+h-s)-u(\cdot,t-s)v(\cdot,t-s)|_{L^1_\sharp}=\zeta(h),$$

which tends to 0 as h tends to 0. Therefore

$$\left|\int_0^t \partial_x \exp(sL)\Big[(uv)(\cdot,t+h-s)-(uv)(\cdot,t-s)\Big]\,ds\right|_{L^2_\sharp}$$
$$\le C\zeta(h)\int_0^T \frac{e^{\rho t}}{t^{3/4}}\,dt.$$

The second term in the right-hand side of (2.11) is handled in a straightforward fashion:

$$\left|\int_t^{t+h} \partial_x \exp(sL)\big[(uv)(\cdot,t+h-s)\big]\,ds\right|_{L^2_\sharp}$$
$$\le C\int_t^{t+h} \frac{e^{\rho t}}{t^{3/4}}\,dt\|u\|_{C^0([0,T];L^2_\sharp)}\|v\|_{C^0([0,T];L^2_\sharp)}.$$

Thus $t\mapsto B_1(u,v)(\cdot,t)$ is continuous from the right on $[0,T)$; a similar argument would show that it is continuous from the left on $(0,T]$: details are left to the reader. ∎

The last information we need is on the following integral equation:

$$y(t)=f(t)+M_1\int_0^t (t-s)^{-3/4}y(s)\,ds, \tag{2.12}$$

and the related integral inequality

$$y(t)\le f(t)+M_1\int_0^t (t-s)^{-3/4}y(s)\,ds, \tag{2.13}$$

where M_1 is a real number.

LEMMA 2.4. *For all $M_1\in\mathbb{R}$ and all f in $L^1_{\rm loc}(\mathbb{R}^+)$, there exists a unique solution $y\in L^1_{\rm loc}(\mathbb{R}^+)$ of (2.12), given by*

$$y(t)=f(t)+\int_0^t K_1(t-s)f(s)\,ds,$$

where the kernel K_1 is locally integrable on $\mathbb{R}^+$ and nonnegative.

If M_1 is a nonnegative number and if f and y are nonnegative functions in $L^1_{\rm loc}(\mathbb{R}^+)$ satisfying (2.13), then for almost every t,

$$y(t)\le f(t)+\int_0^t K_1(t-s)f(s)\,ds.$$

Proof. The kernel K_1 will be obtained explicitly. Denoting by Γ the Euler function, we define for all σ of positive real part a locally integrable function χ^σ over $\mathbb{R}$ by

$$\chi^\sigma=\begin{cases}\Gamma(\sigma)^{-1}x^{\sigma-1} & \text{if } x>0,\\ 0 & \text{otherwise.}\end{cases}$$

It is classical that χ^σ can be extended as an entire function from $\mathbb{C}$ to $\mathcal{D}'(\mathbb{R})$ and that χ^0 can be identified with the Dirac mass δ at 0; see, for instance, [2, chapter III, page 72], where the indices are shifted by 1 relatively to the notation used here. Moreover, we have the following convolution identity $\forall\sigma$ and τ in $\mathbb{C}$:

$$\chi^\sigma * \chi^\tau = \chi^{\sigma+\tau}. \tag{2.14}$$

Equation (2.12) can be solved easily if y and f are extended by 0 for $t \le 0$: it is rewritten as the convolution equation

$$y = f + M\chi^{1/4} * y, \tag{2.15}$$

where $M = M_1\Gamma(1/4)^{-1}$. However, the fundamental solution of (2.15) is

$$\delta + \sum_{j=1}^{\infty} M^j \big(\chi^{1/4}\big)^j = \delta + \sum_{j=1}^{\infty} M^j \chi^{j/4},$$

thanks to (2.14). There remains to check that the kernel

$$K_1(t) = \sum_{j=1}^{\infty} M^j \chi^{j/4}(t)$$

is locally integrable on $\mathbb{R}$. But this kernel contains only three unbounded terms in a neighborhood of 0: those relative to the indices $j = 1$ to 3; it also converges nicely: a direct application of the asymptotic formula for the Γ function yields

$$\frac{\Gamma((j+5)/4)}{\Gamma((j+4)/4)} \sim \big((j+1)/4\big)^{1/4}.$$

Therefore the series

$$K_1(t) - \sum_{j=1}^{3} \frac{M^j t^{(j-4)/4}}{\Gamma(j/4)} = \sum_{j=4}^{\infty} \frac{M^j t^{(j-4)/4}}{\Gamma(j/4)}$$

converges $\forall t > 0$ and its sum is analytical with respect to the variable $t^{1/4}$. The uniqueness of the solution of (2.12) is a consequence of general theorems on the convolution algebra of distributions on $\mathbb{R}$ with support bounded on the left.

Consider the integral inequality

$$y(t) \le f(t) + M_1 \int_0^t (t-s)^{-3/4} y(s)\, ds;$$

here we assume that y and f are nonnegative. Thus we obtain immediately by substituting the inequality satisfied by y in the right-hand side of (2.13):

$$y \le \sum_{j=0}^{n} M^j \chi^{j/4} * f + M^{n+1}\chi^{(n+1)/4} * y.$$

As n tends to infinity, the last term of the right-hand side of this inequality tends to 0 as n tends to infinity, and (2.13) holds in the limit. □

With these preparations, we can pass now to local existence: we transform (1.4) into an integral equation, which can be written as

$$u(\cdot,t) = \exp\bigl(tL\bigr)\phi - \int_0^t \exp\bigl((t-s)L\bigr)\frac{\bigl(u_x(\cdot,s)\bigr)^2}{2}\,ds. \tag{2.16}$$

LEMMA 2.5. *Assume that ϕ belongs to $H^1_\sharp$. For all $R > |\phi|_{H^1_\sharp}$, there exists $\tau > 0$ such that* (2.16) *possesses a unique solution u belonging to $C^0\bigl([0,\tau];H^1_\sharp\bigr)$ which satisfies*

$$\forall t \in [0,\tau], \quad |u(\cdot,t) - \phi|_{H^1_\sharp} \le R. \tag{2.17}$$

Proof. The proof is basically a straightforward application of the strict contraction theorem. Define a mapping $\mathcal{T}$ on $C^0\bigl([0,\tau];H^1_\sharp\bigr)$ by

$$(\mathcal{T}u)(\cdot,t) = \exp(tL)\phi - B_0\bigl(u_x,u_x\bigr)/2. \tag{2.18}$$

Thanks to Lemmas 2.2 and 2.3, if u belongs to the ball $\mathcal{B}_R$ of radius R about 0 in $C^0\bigl([0,\tau];H^1_\sharp\bigr)$, then $\mathcal{T}u$ belongs to $C^0\bigl([0,\tau];H^1_\sharp\bigr)$ and

$$|\mathcal{T}u(\cdot,t)|_{L^2_\sharp} \le e^{\sigma t}|\phi|_{L^2_\sharp} + \frac{CR^2}{2}\int_0^t \frac{e^{\rho s}}{s^{1/4}}\,ds,$$
$$|\partial_x\mathcal{T}u(\cdot,t)|_{L^2_\sharp} \le e^{\sigma t}|\phi_x|_{L^2_\sharp} + \frac{CR^2}{2}\int_0^t \frac{e^{\rho s}}{s^{3/4}}\,ds.$$

Moreover, if u^1 and u^2 belong to $\mathcal{B}_R$, we have

$$(\mathcal{T}u^1 - \mathcal{T}u^2) = -B_0\bigl(u^1_x - u^2_x, u^1_x + u^2_x\bigr)/2,$$

so that

$$\bigl|\bigl(\mathcal{T}u^1 - \mathcal{T}u^2\bigr)(\cdot,t)\bigr|_{L^2_\sharp} \le CR|u^1-u^2|_{C^0([0,\tau];H^1_\sharp)}\int_0^t \frac{e^{\rho s}}{s^{1/4}}\,ds,$$
$$\bigl|\bigl(\partial_x\mathcal{T}u^1 - \partial_x\mathcal{T}u^2\bigr)(\cdot,t)\bigr|_{L^2_\sharp} \le CR|u^1-u^2|_{C^0([0,\tau];H^1_\sharp)}\int_0^t \frac{e^{\rho s}}{s^{3/4}}\,ds.$$

Given $R > |\phi|_{H^1_\sharp}$, we can choose τ such that

$$\begin{aligned}&\left(e^{\sigma\tau}|\phi|_{L^2_\sharp} + \frac{CR^2}{2}\int_0^\tau \frac{e^{\rho s}}{s^{1/4}}\,ds\right)^2\\ &\quad + \left(e^{\sigma\tau}|\phi_x|_{L^2_\sharp} + \frac{CR^2}{2}\int_0^\tau \frac{e^{\rho s}}{s^{3/4}}\,ds\right)^2 \le R^2\end{aligned} \tag{2.19}$$

and

$$C^2R^2\left(\int_0^\tau \frac{e^{\rho s}}{s^{1/4}}\,ds\right)^2 + \left(\int_0^\tau \frac{e^{\rho s}}{s^{3/4}}\,ds\right)^2 < 1. \tag{2.20}$$

Condition (2.19) ensures that $\mathcal{T}$ maps $\mathcal{B}_R$ to itself, and condition (2.20) implies that $\mathcal{T}$ is a strict contraction over $\mathcal{B}_R$. Thus, by the strict contraction principle, there is a unique $u \in \mathcal{B}_R$ which satisfies (2.16). The lemma is proved. ∎

If ϕ is smoother, we have a stronger result.

LEMMA 2.6. *Assume that ϕ belongs to $H^2_\sharp$; then the solution defined on $[0,\tau]$ at Lemma 2.5 belongs to $C^0([0,\tau];H^2_\sharp)$. Moreover, u_t belongs to $C^0\big([0,\tau];L^2_\sharp\big)$ and u solves (1.4).*

Proof. Formally, the second derivative $w = u_{xx}$ satisfies the integral equation

$$w = \exp\big(tL\big)\phi_{xx} - \int_0^t \partial_x \exp\big((t-s)L\big)\big[u_x(\cdot,s)w(\cdot,s)\big]\,ds. \tag{2.21}$$

If we prove that (2.21) possesses a unique solution, then a classical argument will imply that w is indeed the second derivative of u with respect to x. However, we may rewrite (2.21) as

$$w = \exp\big(tL\big)\phi_{xx} - B_1(w,u_x). \tag{2.22}$$

If we define a linear operator $\mathcal{U}$ by

$$\mathcal{U}w = -B_1(w,u_x),$$

and a function $\omega \in C^0([0,\tau];L^2_\sharp)$ by

$$\omega(\cdot,t) = \exp\big(tL\big)\phi_{xx},$$

we solve formally (2.22) by

$$w = \sum_{j=0}^{\infty} \mathcal{U}^j\omega.$$

Estimate (2.10) and Lemma 2.4 imply that the series on the right-hand side of the above equation converges in $C^0([0,\tau];L^2_\sharp)$; the uniqueness is an immediate consequence of the Gronwall-type estimate obtained at Lemma 2.4 applied to the inequality

$$\big|\big(w^1-w^2\big)(\cdot,t)\big|_{L^2_\sharp} \le \int_0^t \frac{Ce^{\rho s}}{s^{3/4}}\,\big|\big(w^1-w^2\big)(\cdot,s)\big|_{L^2_\sharp}\,ds.$$

The classical argument which ensures that the function w obtained as a solution of (2.22) is the second derivative of u with respect to x can be sketched as follows: we define

$$u_h(x,t) = u(x+h,t), \quad \phi_h(x) = \phi(x+h);$$

then, subtracting (2.16) translated by h from (2.16), and differentiating with respect to x, we obtain

$$u_{h,x} - u_x = e^{tL}\big(\phi_{h,x}-\phi_x\big) - B_1\big(u_{h,x}-u_x,(u_{h,x}+u_x)/2\big). \tag{2.23}$$

The triangle inequality implies then that

$$\begin{aligned}&\big|\big(u_{h,x}-u_x\big)(\cdot,t)\big|_{L^2_\sharp}\\ &\quad\le e^{\sigma t}\,|\phi_{h,x}-\phi_x|_{L^2_\sharp} + CR\int_0^t \big|\big(u_{h,x}-u_x\big)(\cdot,s)\big|_{L^2_\sharp}\,\frac{e^{\rho(t-s)}}{(t-s)^{3/4}}\,ds.\end{aligned}$$

If we let $M_1 = RCe^{\rho\tau}$, then we can apply the Gronwall-type estimate of Lemma 2.4, and we find that

$$\big|\big(u_{h,x} - u_x\big)(\cdot,t)\big|_{L^2_\sharp} \le \left(e^{\sigma t} + \int_0^t K_1(t-s)e^{\sigma s}\,ds\right)|\phi_{h,x} - \phi_x|_{L^2_\sharp}\,.$$

Under our assumptions on ϕ,

$$|\phi_{h,x} - \phi_x|_{L^2_\sharp} \le h\,|\phi_{xx}|_{L^2_\sharp}\,,$$

so that, uniformly on $[0,\tau]$,

$$\big|\big(u_{h,x} - u_x\big)(\cdot,t)\big|_{L^2_\sharp} \le C'h. \tag{2.24}$$

We subtract from (2.23) divided by h (2.21), and we get

$$z_h(\cdot,t) = e^{tL}\psi_h - B_1\big(z_h,u_x\big) - B_1\big(u_{h,x} - u_x, (u_{h,x} - u_x)/(2h)\big),$$

where we have used the notations

$$z_h = \frac{u_{x,h} - u_x}{h} - w, \quad \psi_h = \frac{\phi_{h,x} - \phi_h}{h} - \phi_{xx}.$$

A new application of the Gronwall-type inequality of Lemma 2.4 enables us to conclude that

$$\lim_{h\to 0}\ \max_{0\le t\le\tau} |z_h(\cdot,t)|_{L^2_\sharp} = 0,$$

and this proves the first statement of the lemma.

Once that u belongs to $C^0\big([0,\tau];H^2_\sharp\big)$, we see that Lu belongs to $C^0\big([0,\tau];L^2_\sharp\big)$, and $\big(u_x\big)^2$ belongs to $C^0\big([0,\tau];H^1_\sharp\big)$ since $H^1_\sharp$ is a multiplication algebra. A standard argument shows that (1.4) holds in the sense of distributions; thanks to the above functional information, its three terms belong to $C^0\big([0,\tau];L^2_\sharp\big)$, and this concludes the proof of the lemma. □

Let us prove now some a priori estimates.

LEMMA 2.7. *Under the assumptions of Lemma* 2.6*, we have the estimates* $\forall t \in [0,\tau]$

$$|u(\cdot,t)|_{H^1_\sharp} \le \Big(2e^{2\rho t}\,|\phi_x|^2_{L^2_\sharp} + |\phi|^2_{L^2_\sharp}\Big)^{1/2}. \tag{2.25}$$

Proof. Thanks to Lemma 2.6, we can differentiate (1.4) with respect to x; denoting by $v = u_x$, we observe immediately that v satisfies the equation

$$\frac{\partial v}{\partial t} = Lv - \frac{\partial_x\big((v)^2\big)}{2}. \tag{2.26}$$

We multiply (2.26) by v, and we integrate; we find that on the interval of existence $[0,\tau]$,

$$\frac{1}{2}\frac{d}{dt}\int_{\mathbb{T}} |v|^2\,dx = \int_{\mathbb{T}} \big(Lv\big)v\,dx - 2\int_{\mathbb{T}} \big(v\big)^2 v_x\,dx.$$

Under our assumptions, the expression $(v)^2 v_x$ is integrable, so that the last term of the above equality integrates to 0 by periodicity; we have the inequality $\forall k \in \mathbb{N}$:

$$g(k) \leq \rho; \tag{2.27}$$

this inequality and (2.26) imply the differential inequality

$$\frac{d}{dt} |v|^2_{L^2_\sharp} \leq 2\rho |v|^2_{L^2_\sharp},$$

which implies by a Gronwall lemma the following inequality:

$$|v|_{L^2_\sharp} \leq e^{\rho t} |\phi_x|_{L^2_\sharp},$$

which is valid on the interval of existence of u. On the other hand, if we integrate (1.1) on one period, we find that $\widehat{u}(0,t)$ is independent of t. Therefore, we have also the inequality

$$\sum_{k\in\mathbb{Z}} |\widehat{u}(k,t)|^2 \leq \left|\widehat{\phi}(0)\right|^2 + \sum_{k\in\mathbb{Z}\setminus\{0\}} |\widehat{v}(k,t)|^2.$$

These two relations imply that

$$|u(\cdot,t)|^2_{H^1_\sharp} \leq \left(2e^{2\rho t} |\phi_x|^2_{L^2_\sharp} + |\phi|^2_{L^2_\sharp}\right).$$

This shows the desired estimate. □

Now we can complete the proof of Theorem 2.1.

Proof of Theorem 2.1. Let R be defined by

$$R = \left(2e^{2\rho T} |\phi_x|^2_{L^2_\sharp} + |\phi|^2_{L^2_\sharp}\right)^{1/2} + 1, \tag{2.28}$$

and let τ be the upper bound of the existence time for a solution of the integral equation (2.16). Lemma 2.5 implies $\tau > 0$. If τ is larger than T, the theorem is proved; assume that $\tau \leq T$; then thanks to Lemma 2.7 we must have estimate (2.25) $\forall t < \tau$. With the help of Lemma 2.6, we have also

$$|u_{xx}(\cdot,t)|_{L^2_\sharp} \leq e^{\sigma t} |\phi_{xx}|_{L^2_\sharp} + CR \int_0^t \frac{e^{\rho(t-s)}}{(t-s)^{3/4}} |u_{xx}(\cdot,s)|_{L^2_\sharp}\, ds$$

which implies the a priori estimate

$$|u_{xx}|_{L^2_\sharp} \leq e^{\sigma t} |\phi_{xx}|_{L^2_\sharp} + \int_0^t K_1(t-s) e^{\sigma s}\, ds\, |\phi_{xx}|_{L^2_\sharp}.$$

In particular, we infer from (1.4) that u is Lipschitz continuous from $[0,\tau)$ to $L^2_\sharp$ and that the Lipschitz constant depends only on the $H^2_\sharp$ norm of the initial data and on T. These observations imply that the solution can be extended up to the time $t = \tau$. Then thanks to (2.25) we have

$$|u(\cdot,\tau)|_{H^1_\sharp} \leq \left(2e^{2\rho\tau} |\phi_x|^2_{L^2_\sharp} + |\phi|^2_{L^2_\sharp}\right)^{1/2} \leq R - 1.$$

Thanks to the local existence results Lemmas 2.5 and 2.6, we can find $\tau_1 > \tau$ such that the solution exists on $[0,\tau_1]$, takes its values in $H^2_\sharp$, and satisfies the estimate (2.25) with τ replaced by τ_1. This contradicts the assumption that τ was the upper bound of the existence times for which (2.25) holds and concludes the proof of the theorem. □

We have obtained an existence result which used as a technical step the assumption that the initial data belong to $H^2_\sharp$; this assumption can be removed, and we have the following proposition.

PROPOSITION 2.8. *Assume that ϕ belongs to $H^1_\sharp$. Then, $\forall T > 0$, there exists a unique solution u of* (1.4) *in the sense of distributions, which belongs to $C^0\big([0,T], H^1_\sharp\big)$ and which satisfies estimate* (2.25) *and the identity*

$$\forall t \in \mathbb{R}^+, \quad \int_{\mathbb{T}} u(x,t)\,dx = \int_{\mathbb{T}} \phi(x)\,dx. \tag{2.29}$$

Proof. We approximate the initial data $\phi \in H^1_\sharp$ by a sequence $\phi_n \in H^2_\sharp$ which converges to ϕ in the $H^1_\sharp$ norm; Theorem 2.1 implies that there is a solution u_n to (1.4) with initial data ϕ_n, and that this solution satisfies $\forall t \in [0,T]$ the estimate

$$\begin{aligned} |u_n(\cdot,t)|_{H^1_\sharp} &\le \Big(2e^{2\rho t}\,|\phi_{n,x}|^2_{L^2_\sharp} + |\phi_n|^2_{L^2_\sharp}\Big)^{1/2} \\ &\le \Big(2e^{2\rho T}\,|\phi_{n,x}|^2_{L^2_\sharp} + |\phi_n|^2_{L^2_\sharp}\Big)^{1/2} = R. \end{aligned} \tag{2.30}$$

We subtract the integral equation for u_m from the integral equation for u_n and we obtain

$$\begin{aligned} u_n - u_m = {}& \exp\big(Lt\big)(\phi_n - \phi_m) \\ & - B_0\big(u_{n,x} - u_{m,x}, (u_{n,x} + u_{m,x})/2\big). \end{aligned}$$

This relation implies the integral inequality

$$\begin{aligned} &\big|\big(u_{n,x} - u_{m,x}\big)(\cdot,t)\big|_{L^2_\sharp} \\ &\quad \le e^{\sigma t}\,|\phi_{n,x} - \phi_{m,x}|_{L^2_\sharp} + \int_0^t \frac{CRe^{\rho(t-s)}}{(t-s)^{3/4}}\,\big|\big(u_{n,x} - u_{m,x}\big)(\cdot,s)\big|_{L^2_\sharp}\,ds, \end{aligned}$$

and arguing as in the proof of Lemma 2.6, we infer that $\big(u_{n,x}\big)$ is a Cauchy sequence in $C^0\big([0,T]; L^2_\sharp\big)$. A similar argument shows that the sequence $(u_n)_n$ is a Cauchy sequence in $C^0\big([0,T]; H^1_\sharp\big)$. The limit of this Cauchy sequence solves (2.16) and satisfies (2.25); it is classical that (1.4) is solved by u in the sense of distributions. Conversely, if u belongs to $C^0\big([0,T]; H^1_\sharp\big)$, and solves (1.4), it also solves (2.16). The uniqueness is proved as follows: let v be another solution of (1.4) which belongs to $C^0\big([0,T]; H^1_\sharp\big)$, and let

$$R' = \max\big(|u|_{C^0([0,T];H^1_\sharp)}, |v|_{C^0([0,T];H^1_\sharp)}\big);$$

we subtract the integral equation for v from the integral equation for u, and we find that

$$\big(u_x - v_x\big)(\cdot,t) = -B_1\big(u_x - v_x, (u_x + v_x)/2\big),$$

and we obtain immediately the estimate

$$\big|\big(u_x - v_x\big)(\cdot,t)\big|_{L^2_\sharp} \le \int_0^t \frac{CR' e^{\rho s}}{s^{3/4}} |(u_x - v_x)(\cdot, t-s)|_{L^2_\sharp}\, ds;$$

thanks to Lemma 2.4, v must be equal to u. The last estimate holds by continuity: we have observed in the proof of Lemma 2.6 that it holds when the initial data belong to $H^2_\sharp$; therefore, by continuity, it still holds when the initial data belong to $H^1_\sharp$. □

3. Continuity with respect to ε. In this section, we prove the continuity result which is the object of this article. Assume therefore that the sequence g^ε of elements of $\mathcal{G}$ converges to g^0 as in (1.3). We denote by L^ε and L^0 the pseudodifferential operators defined by (1.2) with g replaced, respectively, by g^ε and g^0, and by M^ε and M^0 the analogous operators defined by (2.6). Then, $\forall \psi \in L^2_\sharp$, $\exp(tL^\varepsilon)\psi$ and $\exp(tL^0)\psi$ converge strongly in $L^2_\sharp$, respectively, to $\exp(tL^0)\psi$ and $\exp(tM^0)\psi$. These convergences are uniform on all compact subsets of $L^2_\sharp$ and on the compact interval $[0,T]$.

This strong convergence, together with one more integral inequality, will suffice to prove the last result presented here.

PROPOSITION 3.1. *Assume that ϕ belongs to $H^1_\sharp$; as ε tends to 0, u^ε tends to u^0 in $C^0\big([0,T]; H^1_\sharp\big) \forall T > 0$.*

Proof. Let R be defined by (2.30). We subtract from the integral equation (2.16) the analogous equation for $\varepsilon = 0$ and we get the identity

$$\begin{aligned} u^\varepsilon_x(\cdot,t) - u^0_x(\cdot,t) &= \big[\exp\big(tL^\varepsilon\big) - \exp\big(tL^0\big)\big]\phi_x \\ &\quad - \int_0^t \partial_x \exp\big((t-s)L^\varepsilon\big)\big[u^\varepsilon_x(\cdot,s) - u^0_x(\cdot,s)\big]\frac{u^\varepsilon_x(\cdot,s) + u^0_x(\cdot,s)}{2}\, ds \\ &\quad - \frac{1}{2}\int_0^t \partial_x\big[\exp\big((t-s)L^\varepsilon\big) - \exp\big((t-s)L^0\big)\big]\big(u^0_x(\cdot,s)\big)^2\, ds. \end{aligned} \tag{3.1}$$

We define $y^\varepsilon(t) = \big|u^\varepsilon_x(\cdot,t) - u^0_x(\cdot,t)\big|_{L^2_\sharp}$, and we observe that

$$\begin{aligned} &\left|\int_0^t \partial_x \exp\big((t-s)L^\varepsilon\big)\big[u^\varepsilon_x(\cdot,s) - u^0_x(\cdot,s)\big]\frac{u^\varepsilon_x(\cdot,s) + u^0_x(\cdot,s)}{2}\, ds\right|_{L^2_\sharp} \\ &\quad \le RCe^{\rho T}\int_0^t \frac{y^\varepsilon(s)\, ds}{(t-s)^{3/4}}; \end{aligned}$$

on the other hand, the second integral in the right-hand side of (3.1) is split into a term integrated from 0 to $t-\alpha$ and a term from $t-a$ to t; the second term is estimated as

$$\begin{aligned} &\left|\int_{t-\alpha}^t \partial_x\big[\exp((t-s)L^\varepsilon) - \exp((t-s)L^0)\big]\big[u^0_x(\cdot,s)^2\big]\, ds\right|_{L^2_\sharp} \\ &\quad \le \int_0^\alpha \frac{2Ce^{\rho s}\, ds}{s^{3/4}} \le C'\alpha^{1/4}. \end{aligned}$$

We use the strong convergence of $\exp(tM^\varepsilon)$ to estimate the first term. Indeed, the set $\big\{u^0_x(\cdot,s) : 0 \le s \le T\big\}$ is compact in $L^2_\sharp$, and therefore, the set $\big\{u^0_x(\cdot,s)^2 : 0 \le s \le T\big\}$ is compact in $L^1_\sharp$; thanks to estimate (2.5), $\forall \alpha > 0$, the set

$$\big\{\partial_x \exp\big(t\nu\partial_{xx}\big)u^0_x(\cdot,s)^2 : 0 \le s \le T, \alpha \le t \le T\big\} = K(\alpha)$$

is compact in $L^2_\sharp$; therefore, writing

$$\begin{aligned}&\big[\exp((t-s)L^\varepsilon)-\exp((t-s)L^0)\big]\big[u^0_x(\cdot,s)^2\big]\\&\quad=\big[\exp((t-s)M^\varepsilon)-\exp((t-s)M^0)\big]\exp\big((t-s)\nu\partial_{xx}\big)\big[u^0_x(\cdot,s)^2\big],\end{aligned}$$

we see that

$$\sup_{\substack{\alpha\le t\le T\\0\le s\le t-\alpha}}\big|\partial_x\big[\exp((t-s)L^\varepsilon)-\exp((t-s)L^0)\big]\big[u^0_x(\cdot,s)^2\big]\big|_{L^2_\sharp}=\zeta(\varepsilon,\alpha)$$

tends to 0 as ε tends to 0. Of course, this convergence is *not* uniform with respect to α. Finally, we let

$$\zeta(\varepsilon)=\big|\big[\exp(tL^\varepsilon)-\exp(tL^0)\big]\phi_x\big|_{L^2_\sharp}$$

which tends to 0 as ε tends to 0, thanks to the strong convergence of $\exp(tL^\varepsilon)$ to $\exp(tL^0)$.

Thus y^ε satisfies the integral inequality

$$y^\varepsilon(t)\le\zeta(\varepsilon)+C'\alpha^{1/4}+T\zeta(\varepsilon,\alpha)+M_1\int_0^t\frac{y^\varepsilon(s)\,ds}{(t-s)^{1/4}},$$

where we have chosen

$$M_1=RCe^{\rho T}.$$

Given $\beta>0$, we may always choose α and then ε_0 such that $\forall\varepsilon\le\varepsilon_0$,

$$\zeta(\varepsilon)+C'\alpha^{1/4}+T\zeta(\varepsilon,\alpha)\le\beta;$$

then, the convergence of y^ε to 0 in $C^0([0,T])$ is an immediate consequence of Lemma 2.4. This proves that u^ε_x converges to u^0_x in $C^0([0,T];L^2_\sharp)$. On the other hand, we have seen in the proof of Lemma 2.7 that the zero Fourier coefficient of $u^\varepsilon(\cdot,t)$ is independent of time; thus the convergence of u^ε_x implies the convergence of u^ε in $C^0([0,T];L^2_\sharp)$, hence its convergence in $C^0([0,T];H^1_\sharp)$. This completes the proof of the result. □

4. Conclusion. The result presented here is rather particular. Let us point out some generalizations one might consider: the nonlinearity u^2_x could be replaced by a more general nonlinearity in one dimension; but if the growth of the nonlinearity is faster than quadratic, the local existence theorem presented here fails; it can probably be cured if one is prepared to work with smoother initial data. If one would like to work in dimension 2 or larger, the estimates obtained at Lemma 2.2 are not strong enough: in particular (2.3) is replaced by an analogous estimate in dimension n with a power $t^{-(n+2)/4}$ instead of $t^{-3/4}$. Once we lose the local integrability of this kernel, we have to change methods to obtain something and probably to work in much smoother spaces.

One may wonder whether it could be possible to show estimates on $\exp(tL^\varepsilon)$ as an operator from L^p to L^p: this is indeed possible; however, the lack of precise information on the behavior of $g(k)-g(k-1)$ as $|k|$ tends to infinity seems to exclude semigroup estimates; however, it is quite possible that there should be an estimate with a kernel including negative fractional powers, which might be sufficient for such purposes. But the proof of continuity with respect to these perturbations seems much more difficult in a general case than in the simple case described here.

Acknowledgments. It is a pleasure to thank Guy Joulin for providing me with this question and for sharing with me enlightening conversations on this subject. One of the anonymous referees was instrumental in enhancing the quality of this article by his or her precise reading and well-founded criticisms; deep thanks are owed to her or him.

REFERENCES

[1] G. Joulin and P. Vidal, *An introduction to the stability of flames, shocks and detonations*, in Hydrodynamics and Nonlinear Instabilities, C. Godrèche and P. Manneville, eds., Cambridge University Press, Cambridge, UK, 1998, pp. 493–673.

[2] L. Hörmander, *The Analysis of Linear Partial Differential Operators* I, Springer-Verlag, Berlin, Heidelberg, 1983.

[3] G. I. Sivashinsky, *Nonlinear analysis of hydrodynamic instability in laminar flames.* I. *Derivation of basic equations*, Acta Astronaut., 4 (1977), pp. 1177–1206.

[4] O. Thual, U. Frisch, and M. Hénon, *Application of pole decomposition to an equation governing the dynamics of wrinkled flame fronts*, J. Physique, 46 (1985), pp. 1485–1494.

SIAM J. MATH. ANAL.
Vol. 32, No. 3, pp. 651–664

RESONANCE POCKETS OF HILL'S EQUATIONS WITH TWO-STEP POTENTIALS*

SHAOBO GAN† AND MEIRONG ZHANG‡

Abstract. In this paper, we use the rotation number approach to study in detail the characteristic values of Hill's equations with two-step periodic potentials. As a result, the global structure of resonance pockets is described completely. The results in this paper show that resonance pockets behave in a sensible and fairly rich way even in this simplest case.

Key words. resonance pocket, Hill's equation, characteristic value, rotation number

AMS subject classifications. 34L40, 34D20, 47A10

PII. S0036141099356842

1. Introduction. In this paper we are concerned with the global structure of resonance pockets of parameterized Hill's equations

$$\ddot{x} + (\lambda + \varepsilon p(t))x = 0, \tag{1.1}$$

where $p(t)$ is a 2π-periodic step potential of two steps. For a general 2π-periodic potential, the resonance region R of (1.1) means the set of those parameters (λ, ε) in the (λ, ε)-plane such that (1.1) admits solutions $x(t)$ which are unbounded. The resonance pockets of (1.1), which will be explained more clearly later, are "compact" or "closed" parts of R.

The resonance region R of (1.1) can be described completely in theory. For any fixed parameter ε, R consists of the complement of all spectrum intervals of (1.1). More precisely, let $q(t)$ be a 2π-periodic potential such that $q \in L^1(0, 2\pi)$. Consider the eigenvalue problem

$$\ddot{x} + (\lambda + q(t))x = 0. \tag{1.2}$$

By Theorem 2.1 of Magnus and Winkler [10] or Theorem 8.1, Chapter III of Hale [6], it is well known that problem (1.2) has a sequence of the periodic eigenvalues

$$\lambda_0^P(q) < \lambda_1^P(q) \le \lambda_2^P(q) < \cdots < \lambda_{2n-1}^P(q) \le \lambda_{2n}^P(q) < \cdots$$

with respect to the periodic boundary conditions (P): $x(0) - x(2\pi) = \dot{x}(0) - \dot{x}(2\pi) = 0$. Meanwhile, problem (1.2) also has a sequence of the antiperiodic eigenvalues

$$\lambda_1^A(q) \le \lambda_2^A(q) < \cdots < \lambda_{2n-1}^A(q) \le \lambda_{2n}^A(q) < \cdots$$

with respect to the antiperiodic boundary conditions (A): $x(0) + x(2\pi) = \dot{x}(0) + \dot{x}(2\pi) = 0$. Let us rewrite them as

$$\underline{\lambda}_n(q) = \lambda_n^A(q) \quad \text{and} \quad \overline{\lambda}_n(q) = \lambda_{n+1}^A(q) \qquad \text{when } n \text{ is odd},$$

*Received by the editors May 26, 1999; accepted for publication (in revised form) August 23, 2000; published electronically October 20, 2000. This project was supported by the National Natural Science Foundation of China and the 973 Project of the Ministry of Science and Technology, China.
http://www.siam.org/journals/sima/32-3/35684.html

†School of Mathematical Sciences, Peking University, Beijing 100871, People's Republic of China, and The Abdus Salam International Centre for Theoretical Physics, P.O. Box 586, 34100 Trieste, Italy (gansb@sxx0.math.pku.edu.cn).

‡Department of Mathematical Sciences,Tsinghua University, Beijing 100084, People's Republic of China (mzhang@math.tsinghua.edu.cn).

$$\underline{\lambda}_n(q) = \lambda_{n-1}^P(q) \quad \text{and} \quad \overline{\lambda}_n(q) = \lambda_n^P(q) \qquad \text{when } n \text{ is even.}$$

These eigenvalues, called characteristic values of (1.2) as a whole [10, p. 12], have the following order (see Theorem 2.1 of [10]):

$$\overline{\lambda}_0(q) < \underline{\lambda}_1(q) \le \overline{\lambda}_1(q) < \cdots < \underline{\lambda}_n(q) \le \overline{\lambda}_n(q) < \cdots.$$

Now the resonance region R of (1.1) is given by

$$R = \bigcup_{n=0}^{\infty} R_n,$$

where

$$R_0 = \{(\lambda, \varepsilon) : \lambda < \overline{\lambda}_0(\varepsilon p)\}, \qquad R_n = \{(\lambda, \varepsilon) : \underline{\lambda}_n(\varepsilon p) < \lambda < \overline{\lambda}_n(\varepsilon p)\},\ n = 1, 2, \ldots.$$

A typical example is the Mathieu case: $p(t) = \cos t$. In this case, $\underline{\lambda}_n(p_\varepsilon) < \overline{\lambda}_n(p_\varepsilon)$ holds for all $\varepsilon \neq 0$, $n \in \mathbb{N}$. Thus each resonance region R_n is like a "tongue" which approaches to the point $((n/2)^2, 0)$ on the λ-axis. These are the so-called Arnold tongues (resonance tongues, instability tongues); see section 25, Chapter 5 of [1] and section III.8 of Hale [6]. However, for the near Mathieu case $p(t) = \cos t + \beta \cos 2t$ or the square wave case $p(t) = \operatorname{sign} \cos t$, besides the resonance tongues, it is also observed that some resonance regions R_n would have some closed subregion, namely, $\underline{\lambda}_n(\varepsilon p) = \overline{\lambda}_n(\varepsilon p)$ for some nonzero parameter ε. These interesting phenomena are called resonance pockets; see [1, 4, 6]. One may find in [3] the historical development of the study for resonance regions of Hill's equations. For resonance tongues of certain nonlinear systems, one can refer to [2, 5, 7, 9, 13]. A geometric explanation using singularity theory to the appearance of resonance pockets is given in [3] and has been developed in [2, 4]. Such an idea is very fruitful in explaining the pockets near the λ-axis. However, so far as we know, the global structure for all resonance pockets are not available even for the simplest case—the square wave case.

Note that the problem of resonance pockets of the Hill's equations is just to study the coexistence problem [10, p. 90] of characteristic values:

$$\overline{\lambda}_n(\varepsilon p) = \underline{\lambda}_n(\varepsilon p). \tag{1.3}$$

Such a coexistence problem for general potentials $p(t)$ is extraordinarily difficult. A preliminary idea is to approximate general potentials by step ones. In doing so, we can give a complete analysis of the simplest case, i.e., the 2π-periodic two-step potentials:

$$p(t) = p_{c_1,c_2,t_1}(t) := \begin{cases} c_1 & \text{if } 0 \le t < t_1, \\ c_2 & \text{if } t_1 \le t < 2\pi, \end{cases} \tag{1.4}$$

where $c_1 \neq c_2$, $0 < t_1 < 2\pi$. Denote $t_2 = 2\pi - t_1$. Our result is the following theorem.

THEOREM 1.1. *Let $p(t)$ be given by* (1.4). *Then the number of resonance pockets in the nth resonance region R_n of* (1.1) *is exactly*

$$N_n = \begin{cases} n-2 & \text{if } \frac{nt_1}{2\pi} \text{ is an integer,} \\ n-1 & \text{if } \frac{nt_1}{2\pi} \text{ is not an integer.} \end{cases}$$

This result shows that the coexistence problem (1.3) and the global structure of the corresponding Hill's equations (1.1) depend on the ratio of $t_1/2\pi$ in a very sensible way, while the global structure of (1.1) behaves in an elegant way for "generic" two-step potentials.

COROLLARY 1.2. *When t_1 in (1.4) is incommensurable with π, i.e., t_1/π is irrational, the nth resonance region R_n of (1.1) contains exactly $n-1$ resonance pockets for each $n \in \mathbb{N}$. Moreover, all of resonance pockets are transversal.*

When the square wave potential $p(t)$ (i.e., $c_1 = -1$, $c_2 = +1$, and $t_1 = t_2 = \pi$) is considered, the structure of resonance pockets behaves as follows.

COROLLARY 1.3. *The number of resonance pockets in the nth resonance region R_n of (1.1) with the square wave potential $p(t)$ is exactly*

$$N_n = \begin{cases} n-2 & \text{if } n \text{ is even,} \\ n-1 & \text{if } n \text{ is odd.} \end{cases}$$

Note that the problem for two-step potentials is not too difficult because (1.1) can be solved using trigonometric functions. In particular, the discriminate of (1.1) can be computed explicitly; cf. (1.5). Now characteristic values can be determined by

$$\begin{aligned} \operatorname{tr} P_\lambda &= 2\cos(t_1\sqrt{\lambda+\varepsilon c_1})\cos(t_2\sqrt{\lambda+\varepsilon c_2}) \\ &\quad - \left(\sqrt{\frac{\lambda+\varepsilon c_1}{\lambda+\varepsilon c_2}} + \sqrt{\frac{\lambda+\varepsilon c_2}{\lambda+\varepsilon c_1}}\right)\sin(t_1\sqrt{\lambda+\varepsilon c_1})\sin(t_2\sqrt{\lambda+\varepsilon c_2}) = \pm 2; \end{aligned} \tag{1.5}$$

cf. Lemma 2.3 and p. 116 of [10]. However, (1.5) is not easily analyzed. Due to the coexistence of characteristic values, there is some difficulty in solving (1.5) even numerically. Because of this reason, we adopt in this paper the rotation number approach to characteristic values [8, 11, 12].

The paper is organized as follows. In section 2, the rotation number approach to characteristic values with general periodic potentials is reviewed. Some results concerning the coexistence and the characterization of characteristic values using the solutions of (2.3) (see next section) are given. These results may be of some independent interest. In section 3, we obtain the coexistence conditions and the equations for characteristic values. The results on resonance pockets are proved in section 4.

2. Rotation number approach to characteristic values. Let $\mathcal{P}$ denote the collection of all 2π-periodic functions $q(t)$ such that $q \in L^1(0, 2\pi)$.

Assume that $q \in \mathcal{P}$ and consider eigenvalue problem (1.2). We intend to use the rotation number function to characterize all characteristic values $\underline{\lambda}_n(q)$ and $\overline{\lambda}_n(q)$. Let $y = -\dot{x}$ in (1.2). Then (1.2) is equivalent to the following linear planar system:

$$\dot{x} = -y, \qquad \dot{y} = (\lambda + q(t))x. \tag{2.1}$$

In the polar coordinates: $x = r\cos\theta$, $y = r\sin\theta$,

$$\dot{r} = (\lambda + q(t) - 1)r\cos\theta\sin\theta, \tag{2.2}$$

$$\dot{\theta} = (\lambda + q(t))\cos^2\theta + \sin^2\theta =: \Xi(t, \theta; \lambda). \tag{2.3}$$

Let $\Theta(t; \theta_0, \lambda)$ be the unique solution of (2.3) satisfying the initial condition: $\Theta(0; \theta_0, \lambda) = \theta_0$. As the vector field $\Xi(t, \theta; \lambda)$ is 2π-periodic in t and is π-periodic in θ, one has

$$\Theta(t + 2m\pi; \theta_0, \lambda) = \Theta(t; \Theta(2m\pi; \theta_0, \lambda), \lambda) \tag{2.4}$$

$$\Theta(t; \theta_0 + n\pi, \lambda) = \Theta(t; \theta_0, \lambda) + n\pi \tag{2.5}$$

for all $t,\ \theta_0,\ \lambda \in \mathbb{R}$ and $m,\ n \in \mathbb{Z}$. Thus the rotation number of (2.3)

$$\rho(\lambda) = \rho(\lambda; q) = \lim_{t\to\infty} \frac{\Theta(t; \theta_0, \lambda) - \theta_0}{t}$$

exists and is independent of θ_0; see Theorem 2.1, Chapter 2 of Hale [6].

The solutions $\Theta(t;\theta_0,\lambda)$ depend continuously on the parameter λ. As $\Xi(t,\theta;\lambda)$ is nondecreasing with respect to λ, then so does $\Theta(t;\theta_0,\lambda)$ according to the comparison theorem. From Corollary 2.1, Chapter 2 of Hale [6], one knows that the rotation number function $\rho(\lambda)$ is continuous and nondecreasing. Furthermore, it can be proved that $\rho(\lambda)=0$ for $\lambda \ll -1$, and $\lim_{\lambda\to+\infty}\rho(\lambda)=+\infty$. Now all characteristic values can be determined using $\rho(\lambda)$.

PROPOSITION 2.1. *$\underline{\lambda}_n(q)=\min\{\lambda\in\mathbb{R}:\rho(\lambda)=n/2\}$ for all $n\in\mathbb{N}$, and $\overline{\lambda}_n(q)=\max\{\lambda\in\mathbb{R}:\rho(\lambda)=n/2\}$ for all $n\in\mathbb{Z}^+$.*

Proof. The relationship between spectrum and rotation number has been well developed in [8, 11, 12]. This characterization of characteristic values using rotation number function is a classical result; cf. Theorems 4.3 and 4.4 of [11]. As a proof is not given in [11], we sketch here, for completeness, the proof based on Theorem 2.1 of [10].

Let P_λ be the Poincaré matrix associated with the system (2.1), i.e.,

$$P_\lambda(x_0,y_0)=(x(2\pi;x_0,y_0,\lambda),y(2\pi;x_0,y_0,\lambda)),$$

where $(x(t;x_0,y_0,\lambda),y(t;x_0,y_0,\lambda))$ is the solution of (2.1) satisfying

$$(x(0;x_0,y_0,\lambda),y(0;x_0,y_0,\lambda))=(x_0,y_0).$$

If $\underline{\lambda}_n(q)\le\lambda\le\overline{\lambda}_n(q)$ for some $n\in\mathbb{N}$, it follows from Theorem 2.1 of [10] that $|\operatorname{tr}P_\lambda|\ge 2$ and P_λ has real eigenvalues $\mu_{1,2}$: $P_\lambda v_i=\mu_i v_i$, $v_i\in\mathbb{R}^2\backslash\{0\}$, $i=1,\ 2$. Let $\theta_i\in\mathbb{R}$ be such that $v_i=r_i(\cos\theta_i,\sin\theta_i)$, $i=1,\ 2$. Then $\Theta(2\pi;\theta_i,\lambda)=\theta_i+k_i\pi$ and $\rho(\lambda)=k_1/2=k_2/2=k/2$, where $k=k_\lambda\in\mathbb{Z}$ for each $\lambda\in[\underline{\lambda}_n(q),\overline{\lambda}_n(q)]$. As $\rho(\lambda)$ is continuous, k_λ is independent of $\lambda\in[\underline{\lambda}_n(q),\overline{\lambda}_n(q)]$. In fact, it can be proved that

$$\rho(\lambda)=n/2 \qquad \text{for all } \ \lambda\in[\underline{\lambda}_n(q),\overline{\lambda}_n(q)]. \tag{2.6}$$

On the other hand, if $\lambda\in(\overline{\lambda}_n(q),\underline{\lambda}_{n+1}(q))$ for some $n\in\mathbb{Z}^+$, then $|\operatorname{tr}P_\lambda|<2$. Therefore eigenvalues $\mu_{1,2}$ of P_λ are on the unit circle: $\mu_1=\bar{\mu}_2=e^{\alpha\sqrt{-1}}$ for some $\alpha=\alpha_\lambda\in\mathbb{R}\backslash\pi\mathbb{Z}$. In this case, one has

$$\rho(\lambda)=\alpha/2\pi \pmod{\mathbb{Z}}\notin\frac{1}{2}\mathbb{Z}. \tag{2.7}$$

Now (2.6) and (2.7) show that $\underline{\lambda}_n(q)$ and $\overline{\lambda}_n(q)$ are the endpoints of the interval $\rho^{-1}(n/2)\subset\mathbb{R}$. □

Let $h:\mathbb{R}\to\mathbb{R}$ be a homeomorphism such that

$$h(\vartheta+n\pi)\equiv h(\vartheta)+n\pi \tag{2.8}$$

for all $\vartheta\in\mathbb{R}$ and all $n\in\mathbb{Z}$. One can define the rotation number of h as

$$\rho(h)=\lim_{m\to\infty}\frac{h^m(\vartheta_0)-\vartheta_0}{2m\pi}$$

(independent of the choice of ϑ_0).

Let $h_\lambda:\mathbb{R}\to\mathbb{R}$ be the Poincaré map of (2.3), i.e., $h_\lambda(\vartheta)=\Theta(2\pi;\vartheta,\lambda)$ for $\vartheta\in\mathbb{R}$. By (2.5), h_λ satisfies (2.8). Now the rotation number $\rho(\lambda)$ is same as $\rho(h_\lambda)$.

PROPOSITION 2.2. *Let h be a homeomorphism of $\mathbb{R}$ satisfying* (2.8) *and n be an integer. Then*

(i) $\rho(h) \geq n/2$ *iff* $\max_{\vartheta\in\mathbb{R}}(h(\vartheta) - (\vartheta + n\pi)) \geq 0$.
(ii) $\rho(h) \leq n/2$ *iff* $\min_{\vartheta\in\mathbb{R}}(h(\vartheta) - (\vartheta + n\pi)) \leq 0$.

Proof. Let us prove (i). Assume that $h(\vartheta_0) \geq \vartheta_0 + n\pi$ for some $\vartheta_0 \in \mathbb{R}$. Using (2.8), it is easy to see that $h^m(\vartheta_0) \geq \vartheta_0 + mn\pi$ for all $m \in \mathbb{N}$. Thus

$$\rho(h) = \lim_{m\to+\infty} \frac{h^m(\vartheta_0) - \vartheta_0}{2m\pi} \geq \frac{n}{2}.$$

Conversely, let $M_0 = \max_{\vartheta\in\mathbb{R}}(h(\vartheta) - (\vartheta + n\pi))$. If $M_0 < 0$, we need to prove that $\rho(h) < n/2$. Notice that

$$h(\vartheta) \leq \vartheta + (n\pi + M_0) \qquad \text{for all } \vartheta \in \mathbb{R}$$

implies that

$$h^m(\vartheta) \leq \vartheta + m(n\pi + M_0)$$

for all $m \in \mathbb{N}$ and all $\vartheta \in \mathbb{R}$. Thus

$$\rho(h) = \lim_{m\to+\infty} \frac{h^m(\vartheta) - \vartheta}{2m\pi} \leq \frac{n}{2} + \frac{M_0}{2\pi} < \frac{n}{2}.$$

Conclusion (ii) can be proved similarly. □

PROPOSITION 2.3. *Let n be an integer. Then the following hold.*
(i) $\lambda = \underline{\lambda}_n(q)$ *iff* $\max_{\theta_0}(\Theta(2\pi;\theta_0,\lambda) - (\theta_0 + n\pi)) = 0$.
(ii) $\lambda = \overline{\lambda}_n(q)$ *iff* $\min_{\theta_0}(\Theta(2\pi;\theta_0,\lambda) - (\theta_0 + n\pi)) = 0$.

Proof. By the comparison theorem for solutions, it can be proved that $\Theta(2\pi;\theta_0,\lambda)$ is strictly increasing with respect to λ. Now the results follow from Propositions 2.1 and 2.2. □

It follows from Proposition 2.3 that the coexistence $\overline{\lambda}_n(q) = \underline{\lambda}_n(q)$ can be described using the solutions $\Theta(2\pi;\theta_0,\lambda)$ in the following way.

PROPOSITION 2.4. $\overline{\lambda}_n(q) = \underline{\lambda}_n(q)$ $(= \lambda)$ *iff* $\Theta(2\pi;\theta_0,\lambda) \equiv \theta_0 + n\pi$ *for all* θ_0.

It follows also from Proposition 2.3 that if $\lambda = \overline{\lambda}_n(q)$ or $\lambda = \underline{\lambda}_n(q)$, then it is necessary that there exists some $\vartheta_0 \in \mathbb{R}$ such that

$$\Theta(2\pi;\vartheta_0,\lambda) = \vartheta_0 + n\pi \qquad \text{and} \qquad \left.\frac{d\Theta(2\pi;\vartheta,\lambda)}{d\vartheta}\right|_{\vartheta=\vartheta_0} = 1. \tag{2.9}$$

We show using the Hamiltonian structure of (2.1) that condition (2.9) is also sufficient for λ to be a characteristic value.

PROPOSITION 2.5. $\lambda = \overline{\lambda}_n(q)$ *or* $\underline{\lambda}_n(q)$ *iff* λ *satisfies* (2.9) *for some* $\vartheta_0 \in \mathbb{R}$. $\theta_0 \in \mathbb{R}$.

Proof. For any fixed $\vartheta \in \mathbb{R}$, let $r = R(t;\vartheta,\lambda)$ and $\theta = \Theta(t;\vartheta,\lambda)$ be the solutions of (2.2) and (2.3) satisfying $R(0;\vartheta,\lambda) = 1$ and $\Theta(0;\vartheta,\lambda) = \vartheta$.

Let $P_\lambda : \mathbb{R}^2 \to \mathbb{R}^2$ be the Poincaré map of (2.1). Then P_λ is area-preserving because (2.1) is a Hamiltonian system. Using the solutions $R(t;\vartheta,\lambda)$ and $\Theta(t;\vartheta,\lambda)$, P_λ is given by

$$P_\lambda(r\cos\vartheta,\, r\sin\vartheta) = rR(2\pi;\vartheta,\lambda)(\cos\Theta(2\pi;\vartheta,\lambda),\, \sin\Theta(2\pi;\vartheta,\lambda)) \tag{2.10}$$

for all $r \in \mathbb{R}$ and all ϑ.

Let ϑ_0 be any fixed real number. For any ϑ_1 near ϑ_0, consider the following sector:

$$S = \{(r\cos\vartheta,\, r\sin\vartheta) \in \mathbb{R}^2 : 0 \le r \le 1,\ \vartheta_0 \le \vartheta \le \vartheta_1\}.$$

Then S has area $\frac{1}{2}(\vartheta_1 - \vartheta_0)$. The image $S' = P_\lambda(S)$ is

$$S' = \{(r'\cos\vartheta',\, r'\sin\vartheta') \in \mathbb{R}^2 :\ 0 \le r' \le R(2\pi; \Theta^{-1}(\vartheta';\lambda),\lambda),\\ \Theta(2\pi;\vartheta_0,\lambda) \le \vartheta' \le \Theta(2\pi;\vartheta_1,\lambda)\},$$

where $\Theta^{-1}(\cdot;\lambda)$ is the inverse of $\Theta(2\pi;\,\cdot,\,\lambda)$. Thus S' has area

$$\frac{1}{2}\int_{\Theta(2\pi;\vartheta_0,\lambda)}^{\Theta(2\pi;\vartheta_1,\lambda)} R^2(2\pi;\Theta^{-1}(\vartheta';\lambda),\lambda)d\vartheta' = \frac{1}{2}\int_{\vartheta_0}^{\vartheta_1} R^2(2\pi;\vartheta,\lambda)\frac{d\Theta(2\pi;\vartheta,\lambda)}{d\vartheta}d\vartheta.$$

As P_λ is area-preserving,

$$\frac{1}{2}(\vartheta_1 - \vartheta_0) \equiv \frac{1}{2}\int_{\vartheta_0}^{\vartheta_1} R^2(2\pi;\vartheta,\lambda)\frac{d\Theta(2\pi;\vartheta,\lambda)}{d\vartheta}d\vartheta.$$

Thus

$$\frac{d\Theta}{d\vartheta}(2\pi;\vartheta,\lambda) \equiv \frac{1}{R^2(2\pi;\vartheta,\lambda)}. \tag{2.11}$$

Assume now that $\vartheta_0 \in \mathbb{R}$ satisfies (2.9). Then $\Theta(2\pi;\vartheta_0,\lambda) = \vartheta_0 + n\pi$. Moreover, by the second equality in (2.9) and by (2.11), $R(2\pi;\vartheta_0,\lambda) = 1$. Now we get from (2.10) that

$$\begin{aligned} P_\lambda(\cos\vartheta_0,\,\sin\vartheta_0) &= R(2\pi;\vartheta_0,\lambda)(\cos\Theta(2\pi;\vartheta_0,\lambda),\,\sin\Theta(2\pi;\vartheta_0,\lambda))\\ &= (\cos(\vartheta_0 + n\pi),\,\sin(\vartheta_0 + n\pi))\\ &= (-1)^n(\cos\vartheta_0,\,\sin\vartheta_0). \end{aligned}$$

This shows that P_λ has a nonzero fixed point $(\cos\vartheta_0,\,\sin\vartheta_0)$ if n is even, which yields a nonzero 2π-periodic solution of (2.1). Thus λ is a periodic eigenvalue of (1.2). The case that n is odd implies that λ is an antiperiodic eigenvalue of (1.2). □

3. Two classes of conditions. Let $q(t) \in \mathcal{P}$ be the 2π-periodic potential given by

$$q(t) = q_{b_1,b_2,t_1}(t) := \begin{cases} b_1 & \text{for } 0 \le t < t_1\ (< 2\pi),\\ b_2 & \text{for } t_1 \le t < 2\pi. \end{cases} \tag{3.1}$$

Denote $t_2 = 2\pi - t_1$. We consider the following linear equation:

$$\ddot{x} + q(t)x = 0,$$

or, its equivalent system

$$\dot{x} = -y, \qquad \dot{y} = q(t)x.$$

As in section 2, let $x = r\cos\theta,\ y = r\sin\theta$. Then θ satisfies

$$\dot{\theta} = q(t)\cos^2\theta + \sin^2\theta =: \Xi(t,\theta). \tag{3.2}$$

Let $\Theta(t;\theta_0)$ be the solution of (3.2) satisfying the initial condition $\Theta(0;\theta_0) = \theta_0$. Denote $\Theta(\theta_0) := \Theta(2\pi;\theta_0)$. For any fixed $n \in \mathbb{N}$, we want to find the explicit conditions on $b_1,\ b_2,\ t_1,\ t_2$ so that

$$\Theta(\theta_0) \equiv \theta_0 + n\pi \qquad \text{for all } \theta_0 \in \mathbb{R}. \tag{3.3}$$

By Proposition 2.4, condition (3.3) is related with the coexistence of characteristic values.

In order to study (3.3), we need not consider the trivial case $b_1 = b_2$. Hence we assume that $b_1 \neq b_2$ in (3.1).

PROPOSITION 3.1. *Condition* (3.3) *holds iff* $b_1,\ b_2,\ t_1,\ t_2$ *satisfy* $b_1 > 0,\ b_2 > 0$, *and*

$$t_1\sqrt{b_1} = k\pi \quad \text{and} \quad t_2\sqrt{b_2} = (n-k)\pi \tag{3.4}$$

for some integer k *with* $0 < k < n$.

Proof. Let $\Theta_1(\theta_0) := \Theta(t_1;\theta_0)$. We have four cases to be discussed.

Case 1. $b_1 = a_1^2 > 0$ and $b_2 = a_2^2 > 0$. Assume that (3.3) holds. In this case, by integrating (3.2) on $[0, t_1]$ and $[t_1, 2\pi]$, respectively, we have the following two equalities:

$$\int_{\theta_0}^{\Theta_1(\theta_0)} \frac{d\theta}{a_1^2\cos^2\theta + \sin^2\theta} = t_1, \tag{3.5}$$

$$\int_{\Theta_1(\theta_0)}^{n\pi+\theta_0} \frac{d\theta}{a_2^2\cos^2\theta + \sin^2\theta} = t_2 \tag{3.6}$$

for all θ_0. Differentiating (3.5) and (3.6) with respect to θ_0, one has

$$\frac{1}{a_1^2\cos^2\theta_0 + \sin^2\theta_0} = \frac{\Theta_1'(\theta_0)}{a_1^2\cos^2\Theta_1(\theta_0) + \sin^2\Theta_1(\theta_0)}, \tag{3.7}$$

$$\frac{1}{a_2^2\cos^2\theta_0 + \sin^2\theta_0} = \frac{\Theta_1'(\theta_0)}{a_2^2\cos^2\Theta_1(\theta_0) + \sin^2\Theta_1(\theta_0)} \tag{3.8}$$

for all $\theta_0 \in \mathbb{R}$. From these we obtain

$$\sin(\Theta_1(\theta_0) - \theta_0)\sin(\Theta_1(\theta_0) + \theta_0) \equiv 0.$$

As $\Theta_1(\theta_0)$ is continuous in θ_0, we have either

$$\Theta_1(\theta_0) - \theta_0 \equiv k\pi \qquad \text{for some } k \in \mathbb{Z} \tag{3.9}$$

or

$$\Theta_1(\theta_0) + \theta_0 \equiv k\pi \qquad \text{for some } k \in \mathbb{Z}. \tag{3.10}$$

If (3.9) holds, then k satisfies $0 < k < n$ because $\theta_0 < \Theta_1(\theta_0) < \theta_0 + n\pi$ in this case. Note that

$$\int_0^{\pi} \frac{d\theta}{a^2\cos^2\theta + \sin^2\theta} = \frac{\pi}{a} \qquad (a > 0).$$

It now follows from (3.5) and (3.6) that

$$a_1t_1 = k\pi \quad \text{and} \quad a_2t_2 = (n-k)\pi \qquad \text{for some } 0 < k < n. \tag{3.11}$$

Conversely, if (3.11) is satisfied for some $0 < k < n$, it is easy to see that $\Theta_1(\theta_0) \equiv \theta_0 + k\pi$ and $\Theta(\theta_0) \equiv \Theta_1(\theta_0) + (n-k)\pi \equiv \theta_0 + n\pi$, i.e., equality (3.3) holds for all θ_0.

Assume now that (3.10) is satisfied. Let $\theta_0 = \ell\pi + \alpha$, where $\ell \in \mathbb{Z}$ and $\alpha \in [-\pi/2, \pi/2)$. Thus, by (3.10), $\Theta_1(\theta_0) = (k-\ell)\pi - \alpha$. It follows from (3.5) that

$$\begin{aligned} t_1 &= \int_{\ell\pi+\alpha}^{(k-\ell)\pi-\alpha} \frac{d\theta}{a_1^2\cos^2\theta + \sin^2\theta} \\ &= \left\{ \int_{\ell\pi+\alpha}^{\ell\pi} + \int_{\ell\pi}^{(k-\ell)\pi} + \int_{(k-\ell)\pi}^{(k-\ell)\pi-\alpha} \right\} \frac{d\theta}{a_1^2\cos^2\theta + \sin^2\theta} \\ &= \frac{(k-2\ell)\pi}{a_1} - 2\int_0^{\alpha} \frac{d\theta}{a_1^2\cos^2\theta + \sin^2\theta} \\ &= \frac{(k-2\ell)\pi}{a_1} - \frac{2}{a_1}\arctan\left(\frac{1}{a_1}\tan\alpha\right). \end{aligned}$$

Namely,

$$k\pi - a_1 t_1 = 2\ell\pi + 2\arctan\left(\frac{1}{a_1}\tan\theta_0\right). \tag{3.12}$$

Note that equality (3.12) cannot hold for all $\theta_0 \in \mathbb{R}$. Thus (3.10) cannot happen in this case.

We remark here that if (3.6) is used, one can obtain

$$a_2 t_2 - (n-k)\pi = 2\ell\pi + 2\arctan\left(\frac{1}{a_2}\tan\theta_0\right). \tag{3.13}$$

This also implies that (3.10) cannot happen in this case.

Case 2. $b_1 \le 0$ and $b_2 = a_2^2 > 0$. As $\Psi(\theta) = b_1\cos^2\theta + \sin^2\theta$ has zeros $\theta = \theta_\pm = \pm\arctan\sqrt{-b_1} + j\pi$, $j \in \mathbb{Z}$, we have $\Theta_1(\theta_\pm) = \theta_\pm$. Let now $\theta_0 = \theta_\pm$ in (3.6). Then

$$t_2 = \int_{\theta_\pm}^{\theta_\pm + n\pi} \frac{d\theta}{a_2^2\cos^2\theta + \sin^2\theta} = \frac{n\pi}{a_2}.$$

Thus $a_2 t_2 = n\pi$. This condition, together with (3.6), implies that $\Theta_1(\theta_0) \equiv \theta_0$ for all θ_0, which is impossible because $\Theta_1(\theta_0) = \Theta(t_1; \theta_0)$ is determined by differential equation

$$\dot\theta = b_1\cos^2\theta + \sin^2\theta, \qquad t \in [0, t_1].$$

Case 3. $b_1 > 0$ and $b_2 \le 0$. As characteristic values are invariant under translations of potentials $q_s(t)$ $(= q(t+s))$, one can transfer this case to Case 2.

Case 4. $b_1 \le 0$ and $b_2 \le 0$. In this case the vector field $\Xi(t,\theta) = q(t)\cos^2\theta + \sin^2\theta \le \Psi(\theta) := -\beta^2\cos^2\theta + \sin^2\theta$, where $\beta = \min\{\sqrt{-b_1}\,,\ \sqrt{-b_2}\,\}$. Thus

$$\dot\theta = \Xi(t,\theta) \le -\beta^2\cos^2\theta + \sin^2\theta = \Psi(\theta).$$

As $\Psi(\theta)$ has zeros $\theta_\pm = \pm\arctan\beta + j\pi$, $j \in \mathbb{Z}$, the comparison theorem shows that $\Theta(2\pi; \theta_\pm) \le \theta_\pm$. As a result, (3.3) does not hold for all θ_0. □

Another class of conditions on $b_1,\ b_2,\ t_1,\ t_2$ is when the following holds:

$$\exists\ \theta_0 \text{ such that } \Theta(\theta_0) = \theta_0 + n\pi \text{ and } \left.\frac{d\Theta(\vartheta)}{d\vartheta}\right|_{\vartheta=\theta_0} = 1. \tag{3.14}$$

By Proposition 2.5, condition (3.14) is related with the determination of characteristic values.

PROPOSITION 3.2. *Condition* (3.14) *is equivalent to either*

$$a_1 \sin\frac{a_1 t_1}{2}\cos\frac{a_2 t_2 - n\pi}{2} + a_2 \cos\frac{a_1 t_1}{2}\sin\frac{a_2 t_2 - n\pi}{2} = 0 \tag{3.15}$$

or

$$a_1 \cos\frac{a_1 t_1}{2}\sin\frac{a_2 t_2 - n\pi}{2} + a_2 \sin\frac{a_1 t_1}{2}\cos\frac{a_2 t_2 - n\pi}{2} = 0, \tag{3.16}$$

where $a_1 = \sqrt{b_1}$ *and* $a_2 = \sqrt{b_2}$.

Proof. We consider the first case that $b_1 = a_1^2 > 0$ and $b_2 = a_2^2 > 0$ in the proof of Proposition 3.1. Note that the equalities (3.5) and (3.6) now read as

$$\int_{\vartheta}^{\Theta_1(\vartheta)} \frac{d\theta}{a_1^2\cos^2\theta + \sin^2\theta} = t_1$$

and

$$\int_{\Theta_1(\vartheta)}^{\Theta(\vartheta)} \frac{d\theta}{a_2^2\cos^2\theta + \sin^2\theta} = t_2$$

for all ϑ. Differentiating these equations with respect to ϑ at $\vartheta = \theta_0$, we can once again obtain equalities (3.7) and (3.8) for this specific θ_0 by simply noticing the conditions in (3.14). Now we can proceed as in the proof of Proposition 3.1 and conclude that either (3.11) holds or both of (3.12) and (3.13) hold for this specific θ_0.

Note that (3.11) is a special case of (3.12) and (3.13) with $\ell = 0$ and $\theta_0 = 0$. Eliminating θ_0 from (3.12) and (3.13), we arrive at

$$a_1 \tan\frac{k\pi - a_1 t_1}{2} = a_2 \tan\frac{a_2 t_2 - (n-k)\pi}{2}. \tag{3.17}_k$$

Observe that if $k' = k + 2$ then $(3.17)_{k'}$ is the same as $(3.17)_k$. Thus $(3.17)_k$ yield actually only two equations:

$$a_1 \tan\frac{a_1 t_1}{2} + a_2 \tan\frac{a_2 t_2 - n\pi}{2} = 0,$$

and

$$a_1 \cot\frac{s a_1 t_1}{2} + a_2 \cot\frac{a_2 t_2 - n\pi}{2} = 0.$$

These are just the conditions (3.15) and (3.16), respectively, which are described in the proposition. The converse can also be proved. These prove the proposition for Case 1.

One can prove in the other cases similarly if the complex cosine and sine functions are used in (3.15) and (3.16). □

Let $q(t) = q_{b_1,b_2,t_1}(t)$ be given by (3.1). It follows from Proposition 3.1 that the coexistence $\overline{\lambda}_n(q_{b_1,b_2,t_1}) = \underline{\lambda}_n(q_{b_1,b_2,t_1})$ $(= \lambda)$ is determined by

$$t_1\sqrt{\lambda + b_1} = k\pi \qquad \text{and} \qquad t_2\sqrt{\lambda + b_2} = (n-k)\pi$$

for some $0 < k < n$. Namely, $b_1,\ b_2,\ t_1$ satisfy

$$H_{n,k}: \quad b_2 - b_1 = \left((n-k)\pi/t_2\right)^2 - \left(k\pi/t_1\right)^2, \qquad 0 < k < n. \tag{3.18}$$

We will see from the next section that these surfaces $H_{n,k}$ in the $(b_1,\ b_2,\ t_1)$-space play a fundamental role in analyzing resonance pockets.

4. Application to resonance pockets. Now we apply the results in section 3 to the resonance pockets of Hill's equations (1.1) with two-step potentials, where $p(t) = p_{c_1,c_2,t_1}(t)$ is given by (1.4). Correspondingly, the parameters (b_1, b_2, t_1) in (3.1) are $(c_1\varepsilon, c_2\varepsilon, t_1)$ in this case.

Fix an integer $n \geq 2$. Starting from $\varepsilon = 0$ where $\underline{\lambda}_n(\varepsilon p) = \overline{\lambda}_n(\varepsilon p) = (n/2)^2$, if $\varepsilon \neq 0$ is such that $(c_1\varepsilon, c_2\varepsilon, t_1)$ hits $H_{n,k}$ for some $0 < k < n$, then one gets a resonance pocket inside R_n of (1.1). Explicitly, $(c_1\varepsilon, c_2\varepsilon, t_1) \in H_{n,k}$ is given by

$$\varepsilon = \varepsilon_{n,k} := \frac{1}{c_2 - c_1}\left(((n-k)\pi/t_2)^2 - (k\pi/t_1)^2\right), \tag{4.1}$$

where $\lambda = \underline{\lambda}_n(\varepsilon p) = \overline{\lambda}_n(\varepsilon p)$ is

$$\lambda = \lambda_{n,k} := (k\pi/t_1)^2 - c_1\varepsilon_{n,k} = \frac{1}{c_2 - c_1}\left(c_2\left((n-k)\pi/t_2\right)^2 - c_1\left(k\pi/t_1\right)^2\right). \tag{4.2}$$

Now we can complete the proof of Theorem 1.1. We need only to analyze (4.1). Note that $\varepsilon_{n,k}$ is decreasing when k runs from 1 to $n-1$. If t_1 is such that $nt_1/2\pi$ is not an integer, then all $\varepsilon_{n,k} \neq 0$ for $k = 1, \ldots, n-1$. Note that $\underline{\lambda}_n(\varepsilon p) = \overline{\lambda}_n(\varepsilon p) = (n/2)^2$ when $\varepsilon = 0$. Thus $\underline{\lambda}_n(\varepsilon p) = \overline{\lambda}_n(\varepsilon p)$ iff $\varepsilon = \varepsilon_{n,k}$, $k = 1, \ldots, n-1$, or $\varepsilon = 0$. As a result, R_n contains exactly $n-1$ pockets. When $nt_1/2\pi = k_0$ is an integer, then $0 < k_0 < n$ and $\varepsilon_{n,k_0} = 0$. As a result, $\underline{\lambda}_n(\varepsilon p) = \overline{\lambda}_n(\varepsilon p)$ iff $\varepsilon = \varepsilon_{n,k}$, $k = 1, \ldots, n-1$. Thus R_n contains exactly $n-2$ pockets. This completes the proof of Theorem 1.1. □

We remark that by Proposition 3.2, characteristic values $\lambda = \underline{\lambda}_n(\varepsilon p)$ and $\lambda = \overline{\lambda}_n(\varepsilon p)$ of (1.1) are determined by

$$\begin{aligned}&\sqrt{\lambda + c_1\varepsilon}\,\sin\frac{t_1\sqrt{\lambda + c_1\varepsilon}}{2}\cos\frac{t_2\sqrt{\lambda + c_2\varepsilon} - n\pi}{2}\\ &+\sqrt{\lambda + c_2\varepsilon}\,\cos\frac{t_1\sqrt{\lambda + c_1\varepsilon}}{2}\sin\frac{t_2\sqrt{\lambda + c_2\varepsilon} - n\pi}{2} = 0,\end{aligned} \tag{4.3}$$

$$\begin{aligned}&\sqrt{\lambda + c_1\varepsilon}\,\cos\frac{t_1\sqrt{\lambda + c_1\varepsilon}}{2}\sin\frac{t_2\sqrt{\lambda + c_2\varepsilon} - n\pi}{2}\\ &+\sqrt{\lambda + c_2\varepsilon}\,\sin\frac{t_1\sqrt{\lambda + c_1\varepsilon}}{2}\cos\frac{t_2\sqrt{\lambda + c_2\varepsilon} - n\pi}{2} = 0;\end{aligned} \tag{4.4}$$

see (3.15) and (3.16).

Let $\lambda = \Lambda_1(\varepsilon)$ and $\lambda = \Lambda_2(\varepsilon)$ be the solutions of (4.3) and (4.4) starting at $\Lambda_1(0) = \Lambda_2(0) = (n/2)^2$, respectively. At $(\lambda, \varepsilon) = (\lambda_{n,k}, \varepsilon_{n,k})$, we have

$$\frac{d\Lambda_1}{d\varepsilon} = -\frac{c_1t_1^3(n-k)^2 + c_2t_2^3k^2}{t_1^3(n-k)^2 + t_2^3k^2}, \tag{4.5}$$

$$\frac{d\Lambda_2}{d\varepsilon} = -\frac{c_1t_1 + c_2t_2}{2\pi}, \tag{4.6}$$

when k is odd, and

$$\frac{d\Lambda_1}{d\varepsilon} = -\frac{c_1t_1 + c_2t_2}{2\pi}, \tag{4.7}$$

$$\frac{d\Lambda_2}{d\varepsilon} = -\frac{c_1t_1^3(n-k)^2 + c_2t_2^3k^2}{t_1^3(n-k)^2 + t_2^3k^2}, \tag{4.8}$$

when k is even. Similarly, at the point $(\lambda, \varepsilon) = ((n/2)^2, 0)$, we get from (4.3) and (4.4) that

$$\frac{d\Lambda_1}{d\varepsilon} = -\frac{c_1(\frac{nt_1}{2} + \sin\frac{nt_1}{2}) + c_2(\frac{nt_2}{2} - \sin\frac{nt_1}{2})}{n\pi}, \tag{4.9}$$

$$\frac{d\Lambda_2}{d\varepsilon} = -\frac{c_1(\frac{nt_1}{2} - \sin\frac{nt_1}{2}) + c_2(\frac{nt_2}{2} + \sin\frac{nt_1}{2})}{n\pi}. \tag{4.10}$$

From (4.5)–(4.10), it is easy to check that

$$\left.\frac{d\Lambda_1}{d\varepsilon}\right|_{\varepsilon=\varepsilon_{n,k}} = \left.\frac{d\Lambda_2}{d\varepsilon}\right|_{\varepsilon=\varepsilon_{n,k}} \iff \varepsilon_{n,k} = 0 \tag{4.11}$$

and

$$\left.\frac{d\Lambda_1}{d\varepsilon}\right|_{\varepsilon=0} = \left.\frac{d\Lambda_2}{d\varepsilon}\right|_{\varepsilon=0} \iff \sin\frac{nt_1}{2} = 0. \tag{4.12}$$

Proof of Corollary 1.2. Assume that t_1 is such that t_1/π is irrational. Then $\varepsilon_{n,k} \neq 0$ and $\sin\frac{nt_1}{2} \neq 0$. By (4.11) and (4.12), we have

$$\frac{d\Lambda_1}{d\varepsilon} \neq \frac{d\Lambda_2}{d\varepsilon}$$

at all $(\lambda, \varepsilon) = (\lambda_{n,k}, \varepsilon_{n,k})$ and at $((n/2)^2, 0)$. This means that all resonance pockets in this case are transversal in the (λ, ε)-plane. □

A typical combinatorics structure in this case is plotted in Figure 4.1. The characteristic values are found by solving (4.3) and (4.4) numerically. We remark that suitable ratios for sizes of resonance pockets need a careful choice of the irrational number $t_1/2\pi$ such that it is badly approximated by rational numbers. In Figure 4.1, one of the pockets in R_5 near the λ-axis is very small and is almost invisible.

Assume now that $t_1/2\pi$ is rational. There are two cases to be discussed. The first one is when $n \in \mathbb{N}$ is such that $nt_1/2\pi$ is not an integer. By Theorem 1.1, the nth resonance region R_n of (1.1) has $n-1$ resonance pockets, which are all transversal by (4.11) and (4.12). The second case is when $nt_1/2\pi = k_0$ is an integer. Then all resonance pockets inside R_n, except the two pockets

$$\left\{(\lambda, \varepsilon) : \underline{\lambda}_n(\varepsilon p) < \lambda < \overline{\lambda}_n(\varepsilon p),\ \varepsilon \in (0, \varepsilon_{n,k_0-1})\right\}$$

and

$$\left\{(\lambda, \varepsilon) : \underline{\lambda}_n(\varepsilon p) < \lambda < \overline{\lambda}_n(\varepsilon p),\ \varepsilon \in (\varepsilon_{n,k_0+1}, 0)\right\},$$

are also transversal.

In particular, for the square wave case, i.e., $c_1 = -1$, $c_2 = +1$ and $t_1 = t_2 = \pi$, we know that all pockets inside R_n are transversal if n is odd, and all pockets, except the two pockets

$$\left\{(\lambda, \varepsilon) : \underline{\lambda}_n(\varepsilon p) < \lambda < \overline{\lambda}_n(\varepsilon p),\ 0 < \varepsilon < n\right\}$$

and

$$\left\{(\lambda, \varepsilon) : \underline{\lambda}_n(\varepsilon p) < \lambda < \overline{\lambda}_n(\varepsilon p),\ -n < \varepsilon < 0\right\},$$

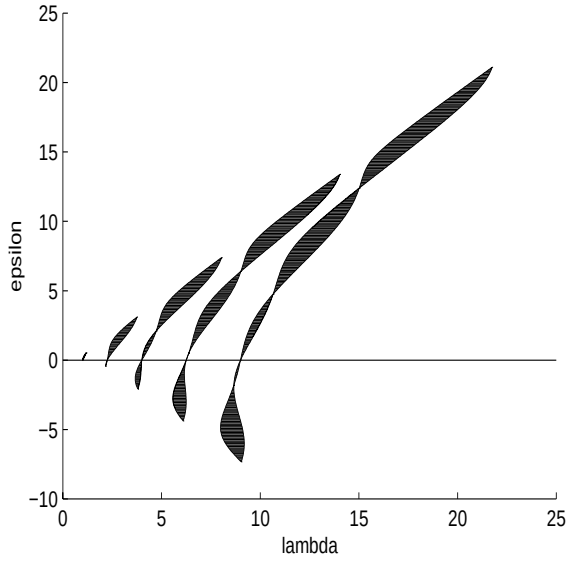


FIG. 4.1. *Resonance pockets for "generic" two-step potentials. Here* $c_1 = -1$, $c_2 = +1$, *and* $t_1 = (\sqrt{5} - 1)\pi$.

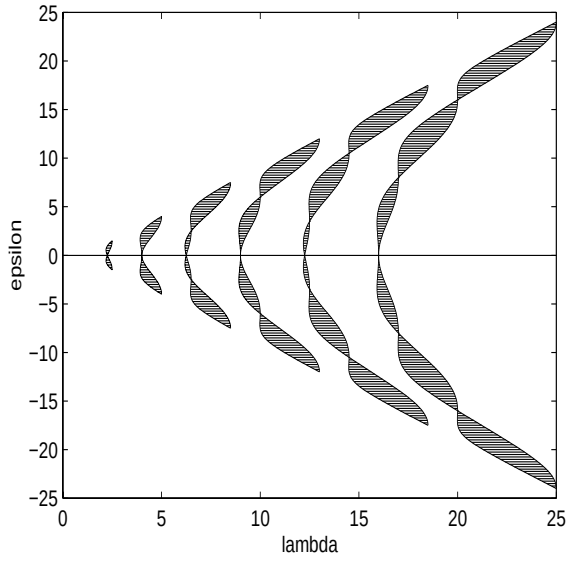


FIG. 4.2. *Resonance pockets for the square wave potential.*

are transversal when n is even. As $-\varepsilon \operatorname{sign} \cos t \equiv \varepsilon \operatorname{sign} \cos(t + \pi)$, the resonance pockets in the square wave case are symmetric with respect to the λ-axis, because characteristic values are invariant when the potentials are translated. This proves Corollary 1.3.

In Figure 4.2, the resonance pockets inside R_n, $n = 3, \ldots, 8$, of the square wave Hill's equations are plotted.

Theorem 1.1 shows that, when $p_\varepsilon(t) = \varepsilon p_{c_1,c_2,t_1}(t)$ is dependent on ε in a linear way, each resonance region of

$$\ddot{x} + (\lambda + p_\varepsilon(t))x = 0 \tag{4.13}$$

contains at most finitely many resonance pockets. However, when general families of two-step potentials

$$p_\varepsilon(t) = p_{b_1(\varepsilon),b_2(\varepsilon),t_1(\varepsilon)}(t)$$

are considered (which depend on ε in a nonlinear way), some resonance regions R_n of (4.13) may contain infinitely many resonance pockets. One example presenting infinitely many resonance pockets inside R_2 is given in [14]. In fact, one can use (3.15), (3.16), and (3.18) to give a global description to all resonance pockets inside all resonance regions R_n of (4.13).

When t_1 is such that t_1/π is irrational, it follows from Corollary 1.2 that all resonance pockets are transversal. This implies that the global structure of resonance pockets of (1.1) is preserved when $p(t)$ has certain kind of smooth perturbations.

Finally, we remark that even for step potentials, the structure of resonance pockets is not easily analyzed. It seems that our approach here is not applicable even to the case $p(t)$ is a three-step potential.

Acknowledgments. The authors thank two anonymous referees for their many suggestions which improve significantly the presentation of the present version. In fact some of their suggestions are actually reflected in the introduction. This revised version was completed when the second author was visiting the Center for Dynamical Systems and Nonlinear Studies, Georgia Institute of Technology. He would like to thank the Center and Professor Shi Jin for their support and their kind hospitality.

REFERENCES

[1] V. I. Arnold, *Mathematical Methods of Classical Mechanics*, Springer-Verlag, New York, 1989.
[2] H. W. Broer, I. Hoveijn, M. van Noort, and G. Vegter, *The inverted pendulum: A singularity theory approach*, J. Differential Equations, 157 (1999), pp. 120–149.
[3] H. W. Broer and M. Levi, *Geometric aspects of stability theory for Hill's equations*, Arch. Ration. Mech. Anal., 131 (1995), pp. 225–240.
[4] H. W. Broer and C. Simó, *Hill's equation with quasi-periodic forcing: Resonance tongues, instability pockets and global phenomena*, Bol. Soc. Brasil. Mat. N.S., 29 (1998), pp. 253–293.
[5] A. M. Davie, *The width of Arnold tongues for the sine circle map*, Nonlinearity, 9 (1996), pp. 421–432.
[6] J. K. Hale, *Ordinary Differential Equations*, 2nd ed., Wiley, New York, 1969.
[7] G. W. Hunt and P. R. Everall, *Arnold tongues and mode-jumping in the supercritical post-buckling of an archetypal elastic structure*, R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci., 455 (1999), pp. 125–140.
[8] R. Johnson and J. Moser, *The rotation number for almost periodic potentials*, Comm. Math. Phys., 84 (1982), pp. 403–438; Erratum, Comm. Math. Phys., 90 (1983), pp. 317–318.
[9] L. B. Jonker, *The scaling of Arnold tongues for differentiable homeomorphisms of the circle*, Comm. Math. Phys., 129 (1990), pp. 1–25.
[10] W. Magnus and S. Winkler, *Hill's Equations*, Wiley, New York, 1966.
[11] J. Moser, *Integrable Hamiltonian Systems and Spectral Theory*, Lezioni Fermiane, Academia Nazionale dei Lincei, Rome, 1983.

[12] J. Moser and J. Pöschel, *An extension of a result by Dinaburg and Sinai on quasi-periodic potentials*, Comment. Math. Helv., 59 (1984), pp. 39–85.
[13] L. Pivka, A. L. Zheleznyak, and L. O. Chua, *Arnold tongues, devil's staircase, and self-similarity in the driven Chua's circuit*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 4 (1994), pp. 1743–1753.
[14] M. Zhang and S. Gan, *Constructing resonance calabashes of Hill's equations using step potentials*, Math. Proc. Cambridge Philos. Soc., 129 (2000), pp. 153–164.

SIAM J. MATH. ANAL.
Vol. 32, No. 3, pp. 665–687

PARTIAL REGULARITY FOR ALMOST MINIMIZERS OF QUASI-CONVEX INTEGRALS*

FRANK DUZAAR†, ANDREAS GASTEL‡, AND JOSEPH F. GROTOWSKI†

Abstract. We consider almost minimizers of variational integrals whose integrands are quasi-convex. Under suitable growth conditions on the integrand and on the function determining the almost minimality, we establish almost everywhere regularity for almost minimizers and obtain results on the regularity of the gradient away from the singular set. We give examples of problems from the calculus of variations whose solutions can be viewed as such almost minimizers.

Key words. quasi convexity, partial regularity, almost minimizers

AMS subject classifications. 49N60, 26B25

PII. S0036141099374536

1. Introduction.

One of the most basic questions in the calculus of variations is that of existence and regularity of minimizers of regular functionals subject to some sort of boundary conditions. To fix ideas we consider a functional

$$\mathcal{F}(u) = \int_U f(x, u, Du)\, dx \tag{1.1}$$

for $x \in U$, a domain in $\mathbb{R}^n$, u mapping U into $\mathbb{R}^N$; then $\mathcal{F}$ is *regular* if $f(x, u, p)$ is convex in p. Appropriate growth conditions on f can be imposed to ensure that the Euler equation corresponding to $\mathcal{F}$ is elliptic, or at least degenerate elliptic; however, even under reasonable assumptions on f, in the case of systems of equations (i.e., $N>1$) one cannot, in general, expect that minimizers of $\mathcal{F}$ will be classical, i.e., C^2-solutions. This was first shown by De Giorgi [DeG]; we refer the reader to [G1, Chapter II.3] for further discussion. It is thus of interest to consider questions of partial regularity. The *regular set* of a solution u is defined by

$$\operatorname{Reg} u = \{x \in U \mid u \text{ is continuous on a neighborhood of } x\}$$

and the *singular set* by

$$\operatorname{Sing} u = U \setminus \operatorname{Reg} u.$$

Partial regularity theory involves estimating the size of $\operatorname{Sing} u$ (i.e., showing that $\operatorname{Sing} u$ has zero n-dimensional Lebesgue measure or better, controlling the Hausdorff dimension of $\operatorname{Sing} u$), and showing higher regularity on $\operatorname{Reg} u$. There is a wealth of literature covering the existence and regularity of minimizers (and, more generally, of stationary points) of regular functionals; we refer the reader to the monographs [G1], [G2], and the literature contained therein.

The condition (for $\mathcal{F}$ to be regular) that the integrand be convex in the gradient is quite restrictive. There are a number of interesting and important problems in the

*Received by the editors October 25, 1999; accepted for publication (in revised form) June 20, 2000; published electronically October 20, 2000.
http://www.siam.org/journals/sima/32-3/37453.html
†Mathematisches Institut der Friedrich-Alexander-Universität Erlangen-Nürnberg, Bismarckstr. 1 1/2, D-91054 Erlangen, Germany (duzaar@mi.uni-erlangen.de, grotow@mi.uni-erlangen.de).
‡Mathematisches Institut der Heinrich-Heine-Universität Düsseldorf, Universitätsstrasse 1, D-40225 Düsseldorf, Germany (gastel@cs.uni-duesseldorf.de).

calculus of variations which are not regular; in addition, weak lower semicontinuity, an essential notion for showing the existence of minimizers, is implied by convexity (in appropriate Sobolev spaces), but not vice versa. This led Morrey to introduce the notion of *quasi convexity* in the paper [M1]; we postpone giving a precise definition until section 2 and simply note here that Morrey showed that, in many circumstances, quasi convexity and weak lower semicontinuity are equivalent, and refer the reader additionally to [Da], [Ba], and [AF] for discussion, literature, and further references.

The first results on partial regularity for minimizers of general quasi-convex integrands were obtained by Evans [Ev]. He considered integrals of the form $\mathcal{F}(u) = \int_U f(Du)\,dx$ and showed, under the principle assumption of uniform strict quasi convexity (see (H2) of the current paper), that a minimizer u of such an $\mathcal{F}$ satisfies $\mathcal{L}^n(\operatorname{Sing} u) = 0$ and that Du is Hölder continuous for all exponents between 0 and 1; see [Ev, section 2] for precise statements. These results were extended independently by Fusco–Hutchinson [FH] and Giaquinta–Modica [GM] to more general functionals of the form (1.1) under assumptions comparable to our (H1)–(H4) and to an additional assumption concerning the Hölder continuity of the integrand $f(x, u, p)$ in x and u; see [FH, section 2] and [GM, Theorem 1.1]. Note in particular that in these results Du is shown to be Hölder continuous for some exponent depending on the Hölder continuity of the integrand f.

In the current paper we wish to consider a more general class of functions than minimizers, namely, *almost minimizers*. Writing $\mathcal{F}(u; D)$ for $\int_D f(x, u, Du)\,dx$, an almost minimizer (at x_0) for $\mathcal{F}$ is a function u for which

$$\mathcal{F}(u; B_\rho(x_0)) \le \mathcal{F}(u + \varphi; B_\rho(x_0)) + \omega(\rho) \int_{B_\rho(x_0)} (1 + |Du|^2 + |D\varphi|^2)dx \tag{1.2}$$

for all suitable test functions φ with $\operatorname{supp}\varphi \subset B_\rho(x_0)$; see Definition 2.1 for a precise statement. Here ω is a real-valued function. Obviously ω identically vanishing corresponds to the case of $\mathcal{F}$-minimizers, and minimal conditions on ω (continuous and nondecreasing at 0 with $\omega(0) = 0$) ensure that the term almost minimizer makes sense. In the next section we impose some additional (mild) conditions on ω and give examples that show that solutions of a number of problems in the calculus of variations (precisely, minimizers subject to certain constraints) are almost minimizers of suitable functionals; hence the notion of an almost minimizer is in fact useful.

A comparable but more restrictive definition of an almost minimizer was given by Anzellotti [An]. In that paper the author shows partial regularity for almost minimizers of the (regular) functional with integrand given by $a^{\alpha\beta}(x)D_\alpha u D_\beta u + g(x)$ for suitably regular $a^{\alpha\beta}$ and g; see [An, Theorem 1.5]. Anzellotti's definition was more restrictive in two respects; he required Hölder continuity for the function ω and required a sharper inequality than (1.2). We also mention that there is another related concept for regular integrands, namely, that of a quasi minimizer (or Q-minimizer); here the right-hand side of (1.2) is replaced by $Q\,\mathcal{F}(u + \varphi; B_\rho(x_0))$ for some constant $Q \ge 1$; see [G1, Chapter IX] for details and further references.

We also note here that there are close ties between the current setting and the study of elliptic parametric variational problems in geometric measure theory. In particular, our notion of an almost minimizer is analogous to Almgren's definition of an $(\mathbf{F}, \varepsilon, \delta)$-minimizer; see [Al, Chapter III]. Indeed our regularity result, Theorem 2.2, is the analogue of Almgren's regularity theorem [Al, Theorem III.3.7] in the current setting; of course [Al, Theorem III.3.7] is broader in scope, and the proof is considerably more involved than the proof of our regularity result. We refer the

reader to [Ev, section 1] for more comments on the connections to geometric measure theory and restrict ourselves here to noting the above-mentioned work of Almgren [Al], as well as the paper of Bombieri [Bo]. The closest analogue of the current paper in the setting of geometric measure theory is the paper [DS], where the authors prove optimal regularity results for almost minimizing rectifiable currents of general elliptic integrands.

The main regularity result of this paper is given in Theorem 2.2. We consider integrals of the form $F(u) = \int_U f(Du)\,dx$ and show, under reasonable conditions on f (the main one being uniform strict quasi convexity) and the function ω, that (F, ω)-minimizers are regular away from a set of zero-measure. In addition we obtain an optimal local modulus of continuity for Du on $\operatorname{Reg} u$. The structure of the proof and the nature of our definition of an almost minimizer enable us to extend this result to families of such integrals. This allows us to obtain, as an easy corollary, partial regularity for *minimizers* of integrals of the form $F(u) = \int_U f(x, Du)\,dx$, where f is quasi-convex, but where we only require a Dini condition (cf. [HW, section 1]) on the continuity of the coefficients in x. In particular, we do not need to assume that the coefficients are Hölder continuous with respect to x, in contrast to the results of [FH] and [GM] (of course, the results there admit u-dependency, in contrast to the current paper). Indeed, even for minimizers of regular integrals of the form $F(u) = \int_U f(x, Du)\,dx$, in the case of systems (i.e., $N > 1$) this appears to be the first time that partial regularity results have been obtained for coefficients which are not Hölder continuous (there are a number of results for scalar valued problems; we mention here specifically [HW] and the recent paper [Ko]).

We wish to briefly comment on our technique. The central idea in our proof is that of A-harmonic approximation, as expressed in Lemma 4.2. This idea, too, has its origins in the field of geometric measure theory, specifically in Simon's proof of the regularity theorem of Allard [A]; see [S1, section 23], and cf. [Bo]. The point here is to show that for $A \in \operatorname{Bil}(\operatorname{Hom}(\mathbb{R}^n, \mathbb{R}^N))$, which is rank-one elliptic, a function which is "approximately A-harmonic," i.e., a function g for which $\int_U A(Dg, D\varphi)\,dx$ is sufficiently small for all test functions φ, lies L^2-close to some A-harmonic function. Lemma 4.2 is due to Duzaar–Steffen (see [DS, Lemma 3.3]). The lemma is also vital to the paper [DG], where the authors give an elementary, self-contained approach to partial regularity for nonlinear elliptic systems of divergence type.

Many of the advantages of the approach of [DG] are relevant in the current paper. In particular we note that the arguments in both papers avoid the technical complications associated with using Gehring's lemma [Ge]; as noted above, in the current setting this is essential to obtaining the optimal modulus of continuity. Furthermore the A-harmonic approximation lemma is the only time where we argue indirectly; hence we keep some control on the sensitivity to the structure constants in our proof.

In section 2 we discuss our assumptions on the integrand f and the function ω and give a number of examples (as discussed above, these are concerned with applications of the partial regularity theorem and with showing that the notion is in fact useful; we also show how the result is optimal in a certain sense). The remainder of the paper is concerned with the proof of the regularity theorem.

We close this section by briefly summarizing the notation we use in this paper. As noted above, we consider a domain $U \subset \mathbb{R}^n$ and maps from U to $\mathbb{R}^N$, where we take $n \geq 2$, $N \geq 1$. For a given set X we denote by $\mathcal{L}^n(X)$ its n-dimensional Lebesgue measure. We write $B_\rho(x_0) = \{x \in \mathbb{R}^n : |x - x_0| < \rho\}$, and further $B_\rho = B_\rho(0)$, $B = B_1$. For bounded $X \subset \mathbb{R}^n$ we denote the average of a given $g \in L^1(X)$ by

$⨍_X g\,dx$, i.e., $⨍_X g\,dx = \frac{1}{\mathcal{L}^n(X)}\int_X g\,dx$. In particular, we write $g_{x_0,\rho} = ⨍_{B_\rho(x_0)} g\,dx$. We let α_n denote the volume of the unit ball in $\mathbb{R}^n$, i.e., $\alpha_n = \mathcal{L}^n(B)$. We write $\mathrm{Bil}(\mathrm{Hom}(\mathbb{R}^n,\mathbb{R}^N))$ for the space of bilinear forms on the space $\mathrm{Hom}(\mathbb{R}^n,\mathbb{R}^N)$ of linear maps from $\mathbb{R}^n$ to $\mathbb{R}^N$.

2. Assumptions, examples, and the partial regularity theorem. We consider a function $\omega : [0,\infty) \to [0,\infty)$, and define

$$\Omega(r) := \left(\int_0^r \frac{\sqrt{\omega(\rho)}}{\rho}\,d\rho\right)^2.$$

We impose the following conditions:

(ω0) ω is nondecreasing;

(ω1) $r \mapsto \omega(r)/r^{2\alpha}$ is nonincreasing for some $\alpha \in (0,1)$;

(ω2) $\omega(r) \le 1$ for all r; and

(ω3) $\Omega(r)$ is finite for some $r > 0$.

Note that all the arguments involving ω in this paper are local in nature; therefore (ω2) is always realizable. In addition (ω3) shows that $\Omega(r)$ is in fact finite for all positive r. Before we discuss some of the consequences of (ω0)–(ω3) we define the central concept of the paper, that of an almost minimizer.

DEFINITION 2.1. *Consider a functional $\mathcal{F}$ defined on $H^{1,2}_{loc}(U,\mathbb{R}^N)$ and $\omega : [0,\infty) \to [0,\infty)$. A function $u \in H^{1,2}_{loc}(U,\mathbb{R}^N)$ is called $(\mathcal{F},\omega)$-minimizing at $x_0 \in U$ if, for all $\rho > 0$ with $B_\rho(x_0) \subset\subset U$, there holds*

$$\mathcal{F}(u;B_\rho(x_0)) \le \mathcal{F}(u+\varphi;B_\rho(x_0)) + \omega(\rho)\int_{B_\rho(x_0)} (1+|Du|^2+|D\varphi|^2)\,dx \tag{2.1}$$

for all $\varphi \in H^{1,2}_0(B_\rho(x_0),\mathbb{R}^N)$.

A function u is $(\mathcal{F},\omega)$-minimizing if u is $(\mathcal{F},\omega)$-minimizing at each $x_0 \in U$.

We now note some less immediate consequences of the above conditions, which we will need in section 5. From (ω1) we see

$$\omega(tr) \le t^{2\alpha}\omega(r) \qquad \text{for } t \ge 1, \tag{2.2}$$

and from the definition of Ω we thus have

$$\Omega(tr) \le t^{2\alpha}\Omega(r) \qquad \text{for } t \ge 1. \tag{2.3}$$

We further have, for $0 < \tau < 1$, $r > 0$, $j \in \mathbb{N} \cup \{0\}$

$$\frac{1}{\alpha}(1-\tau^\alpha)\sqrt{\omega(\tau^j r)} = \frac{\sqrt{\omega(\tau^j r)}}{(\tau^j r)^\alpha}\int_{\tau^{j+1}r}^{\tau^j r} \rho^{\alpha-1}\,d\rho \le \int_{\tau^{j+1}r}^{\tau^j r} \frac{\sqrt{\omega(\rho)}}{\rho}\,d\rho. \tag{2.4}$$

This estimate has two useful consequences. We first note

$$\sum_{j=0}^{\infty} \sqrt{\omega(\tau^j r)} \le \frac{\alpha}{1-\tau^\alpha}\sqrt{\Omega(r)}. \tag{2.5}$$

In addition we see

$$\omega(r) \le \Omega(r) \tag{2.6}$$

for all $r > 0$. We note further that ($\omega 0$) and ($\omega 1$) imply continuity of ω at 0, as well as $\omega(0) = 0$.

We now discuss our assumptions on the functional in question. We consider functionals of the form

$$F(u) := \int_U f(Du)\, dx,$$

where U is a domain in $\mathbb{R}^n$, and $f : \mathrm{Hom}(\mathbb{R}^n, \mathbb{R}^N) \to \mathbb{R}$ satisfies the following conditions:

(H1) there exist positive constants c_1 and c_2 such that, for all $p \in \mathrm{Hom}(\mathbb{R}^n, \mathbb{R}^N)$,

$$c_1^{-1}|p|^2 - c_2 \le f(p) \le c_1|p|^2 + c_2;$$

(H2) the function f is *(uniformly) strictly quasi-convex*, i.e., there exists $\lambda > 0$ such that for all $B_\rho(x_0) \subset\subset U$, $p \in \mathrm{Hom}(\mathbb{R}^n, \mathbb{R}^N)$, $\varphi \in C_0^1(B_\rho(x_0), \mathbb{R}^N)$ there holds

$$\int_{B_\rho(x_0)} \Big(f(p + D\varphi) - f(p) \Big)\, dx \ge \lambda \int_{B_\rho(x_0)} |D\varphi|^2\, dx;$$

(H3) the function f is C^2 and there exists a nonnegative constant L such that for all $p \in \mathrm{Hom}(\mathbb{R}^n, \mathbb{R}^N)$ there holds $|D^2 f(p)| \le L$.

Note that the upper bound in (H1) follows from (H3), and the lower bound is only useful for questions of existence; cf. [M2, 4.4.7], [Ev, p. 228]. We include the condition here largely for completeness in the examples which follow.

Condition (H2) implies the *Legendre–Hadamard condition*; see [M2, 4.4.3, 4.4.1] or [Fe, 5.1.10], i.e.,

$$\sum_{i,j=1}^{N} \sum_{\alpha,\beta=1}^{n} \frac{\partial^2 f}{\partial p_\alpha^i \partial p_\beta^j}(p) \xi^i \xi^j \eta_\alpha \eta_\beta \ge \lambda |\xi|^2 |\eta|^2 \tag{2.7}$$

for all $p \in \mathrm{Hom}(\mathbb{R}^n, \mathbb{R}^N)$, $\xi \in \mathbb{R}^N$, and $\eta \in \mathbb{R}^n$.

From condition (H3) we have

$$|Df(p) - Df(\tilde{p})| \le L|p - \tilde{p}|\,; \tag{2.8}$$

this condition also implies the existence of a modulus of continuity of $D^2 f$, more precisely of a family of monotone nondecreasing, concave functions $\nu(M, \cdot) : [0, \infty) \to [0, \infty)$ for $M > 0$ satisfying $\nu(M, 0) = 0$ and

$$|D^2 f(p) - D^2 f(\tilde{p})| \le \nu(M, |p - \tilde{p}|^2) \tag{2.9}$$

for all $p, \tilde{p} \in \mathrm{Hom}(\mathbb{R}^n, \mathbb{R}^N)$ with $|p| \le M$.

For the proof of our main theorem we will initially strengthen (H3) by further imposing

(H4) $D^2 f$ is uniformly continuous.

In conjunction with (H3) this leads to the existence of a monotone nondecreasing, concave function $\nu : [0, \infty) \to [0, \infty)$ satisfying $\nu(0) = 0$ and

$$|D^2 f(p) - D^2 f(\tilde{p})| \le \nu(|p - \tilde{p}|^2) \tag{2.10}$$

for all $p, \tilde{p} \in \mathrm{Hom}(\mathbb{R}^n, \mathbb{R}^N)$. At the end of the paper (Corollary 5.3) we show how the arguments can be modified to remove (H4).

We are now in a position to state our main result.

THEOREM 2.2. *On a domain $U \subseteq \mathbb{R}^n$ consider a function ω satisfying ($\omega 0$)–($\omega 3$), and a function f which satisfies* (H2) *and* (H3). *Let F be the functional on $H^{1,2}(U,\mathbb{R}^N)$ given by $F(u) = \int_U f(Du)\,dx$. Let $u \in H^{1,2}(U,\mathbb{R}^N)$ be (F,ω)-minimizing on U. Then there exists a relatively closed subset of U,* $\operatorname{Sing} u$, *such that*

$$u \in C^1(U \setminus \operatorname{Sing} u)\,.$$

Further $\operatorname{Sing} u \subseteq \Sigma_1 \cup \Sigma_2$, *where here*

$$\Sigma_1 = \left\{ x_0 \in U \ : \ \liminf_{\rho\to 0^+} \fint_{B_\rho(x_0)} |Du - (Du)_{x_0,\rho}|^2\,dx > 0 \right\}, \quad \textit{and}$$

$$\Sigma_2 = \left\{ x_0 \in U \ : \ \sup_{\rho>0} |(Du)_{x_0,\rho}| = \infty \right\}\ ;$$

in particular $\mathcal{L}^n(\operatorname{Sing} u) = 0$.

In addition, in a neighborhood of any $x_0 \in U\setminus \operatorname{Sing} u$ and for any β with $\alpha < \beta < 1$, Du has a modulus of continuity given by

$$\mu(r) = c\left(r^\beta + \sqrt{\Omega(r)}\right)\,,$$

where c is a constant depending only on $\limsup_{\rho\to 0} |(Du)_{x_0,\rho}|$, *on β, on the dimensions n and N, on the structural parameters λ, L, and α, and on the functions $\omega(\cdot)$ and $\nu(\cdot)$.*

With a view to applications (see, in particular, Example 1 below) we are also interested in being able to consider a different functional at each point, i.e., a functional of the form

$$F_{x_0}(u) := \int_U f_{x_0}(Du)\,dx$$

for $x_0 \in U$. Given a family of such functionals, the analogues of (H2) and (H3) are

(h2) the functions f_{x_0} are *uniformly strictly quasi-convex*, i.e., there exists $\lambda > 0$ such that for all $B_\rho(x_0) \subset\subset U$, $p \in \operatorname{Hom}(\mathbb{R}^n,\mathbb{R}^N)$, $\varphi \in C_0^1(B_\rho(x_0),\mathbb{R}^n)$ there holds

$$\int_{B_\rho(x_0)} \Big(f_{x_0}(p + D\varphi) - f_{x_0}(p) \Big)\,dx \geq \lambda \int_{B_\rho(x_0)} |D\varphi|^2\,dx;$$

(h3) the functions f_{x_0} are C^2 and there exists $L \geq 0$ such that for all $p \in \operatorname{Hom}(\mathbb{R}^n,\mathbb{R}^N)$ and $x_0 \in U$ there holds $|D^2 f_{x_0}(p)| \leq L$.

Just as we imposed the additional condition (H4) to obtain a uniform modulus of continuity above, we will have occasion to require that

(h4) the second derivatives $D^2 f_{x_0}$ admit a *uniform* modulus of continuity, i.e., there exists a monotone nondecreasing, concave function $\nu : [0,\infty) \to [0,\infty)$ satisfying $\nu(0) = 0$ and

$$|D^2 f_{x_0}(p) - D^2 f_{x_0}(\tilde{p})| \leq \nu(|p - \tilde{p}|^2) \tag{2.11}$$

for all $p, \tilde{p} \in \operatorname{Hom}(\mathbb{R}^n,\mathbb{R}^N)$ and $x_0 \in U$.

We can now state the regularity result for families of functionals: the proof follows exactly the same lines as the proof of Theorem 2.2.

COROLLARY 2.3. *The conclusion also holds when* $\{F_{x_0}\}_{x_0\in U}$ *is a family of functionals arising from functions* $\{f_{x_0}\}_{x_0\in U}$ *satisfying* (h2), (h3), *and* (h4), ω *is as above, and* $u \in H^{1,2}(U,\mathbb{R}^N)$ *is* (F_{x_0},ω)*-minimizing at each* $x_0 \in U$.

We now give a few examples of almost minima and applications of the partial regularity result.

Example 1. Consider u minimizing a functional of the form

$$G(u) := \int_U g(x, Du)\,dx\,,$$

where here the frozen coefficients

$$g_{x_0}(p) := g(x_0, p)$$

satisfy (h2), (h3), and (h4), and in addition

$$|g(x,p) - g(\tilde{x},p)| \le \omega(|x-\tilde{x}|)(1+|p|^2) \tag{2.12}$$

for all $x, \tilde{x} \in U$ and all $p \in \mathrm{Hom}(\mathbb{R}^n,\mathbb{R}^N)$ for some ω satisfying (ω1)–(ω3). Writing

$$G_{x_0}(u) := \int_U g_{x_0}(Du)\,dx = \int_U g(x_0, Du)\,dx,$$

we have that u is (G_{x_0},ω)-minimizing at each $x_0 \in U$.

Example 2 (solutions of an obstacle problem). We wish to minimize $\int_U |Dv|^2\,dx$ amongst all functions $v \in H_0^{1,2}(U,\mathbb{R}^N)$ satisfying

$$v^i \ge \psi^i, \qquad (i = 1,\dots,N),$$

where the given functions ψ^i are nonpositive on ∂U and in the class $C^{1,\alpha}$. In order to see that a minimizer u is an almost minimizer of the Dirichlet integral with $\omega(\rho) = c\rho^{2\alpha}$ (for a positive constant c), we argue as follows. (Note that this example is essentially the same as [An, Example 3.2], but for completeness we repeat the arguments here.)

Fix $B_\rho(x_0) \subset\subset U$ and let $h : B_\rho(x_0) \to \mathbb{R}^N$ be the (vector-valued) harmonic function coinciding with u on $\partial B_\rho(x_0)$. Since h is harmonic (and hence minimizing), we have

$$\begin{aligned}\int_{B_\rho(x_0)} |Du|^2\,dx &= \int_{B_\rho(x_0)} |Dh|^2\,dx + \int_{B_\rho(x_0)} |D(u-h)|^2\,dx \\ &\le \int_{B_\rho(x_0)} |D(u+\varphi)|^2\,dx + \int_{B_\rho(x_0)} |D(u-h)|^2\,dx\end{aligned} \tag{2.13}$$

for all $\varphi \in H_0^{1,2}(B_\rho(x_0),\mathbb{R}^N)$. On the other hand, the harmonicity of h and the minimality of u also imply

$$\begin{aligned}&\int_{B_\rho(x_0)} D(u-h)\cdot D(u-v)\,dx = \int_{B_\rho(x_0)} Du\cdot D(u-v)\,dx \\ &= \frac{1}{2}\frac{d}{dt}\bigg|_{t=0^+}\left[\int_{B_\rho(x_0)} |Du|^2\,dx - \int_{B_\rho(x_0)} |(1-t)Du + tDv|^2\,dx\right] \le 0\end{aligned}$$

for all $v \in H^{1,2}(B_\rho(x_0), \mathbb{R}^N)$ with $v = u$ on $\partial B_\rho(x_0)$ and $v^i \geq \psi^i$. We set $v^i = h^i \vee \psi^i = \max\{h^i, \psi^i\}$ for $i = 1, \ldots, N$ and infer

$$\int_{B_\rho(x_0)} D(u-h) \cdot D(u - h \vee \psi)\, dx \leq 0;$$

hence

$$\begin{aligned}\int_{B_\rho(x_0)} |D(u-h)|^2\, dx &\leq \int_{B_\rho(x_0)} D(u-h) \cdot D(h \vee \psi - h)\, dx \\ &\leq \frac{1}{2}\int_{B_\rho(x_0)} |D(u-h)|^2\, dx + \frac{1}{2}\int_{B_\rho(x_0)} |D(h \vee \psi - h)|^2\, dx\end{aligned}$$

and therefore

$$\int_{B_\rho(x_0)} |D(u-h)|^2\, dx \leq \int_{B_\rho(x_0)} |D(h \vee \psi - h)|^2\, dx. \tag{2.14}$$

The last integral can be estimated by $c\rho^{n+2\alpha}$, as can be seen by the inequality

$$\begin{aligned}&\int_{B_\rho(x_0)} |D(h^i \vee \psi^i - h^i)|^2\, dx \\ &= \int_{B_\rho(x_0)} (D(h^i \vee \psi^i) - (D\psi^i)_{x_0,\rho}) \cdot D(h^i \vee \psi^i - h^i)\, dx \\ &\leq \int_{B_\rho(x_0)} |D(h^i \vee \psi^i) - (D\psi^i)_{x_0,\rho}|\, |D(h^i \vee \psi^i - h^i)|\, dx \\ &= \int_{\{h^i \leq \psi^i\}} |D\psi^i - (D\psi^i)_{x_0,\rho}|\, |D(\psi^i - h^i)|\, dx \\ &\leq \frac{1}{2}\int_{B_\rho(x_0)} |D\psi^i - (D\psi^i)_{x_0,\rho}|^2\, dx + \frac{1}{2}\int_{B_\rho(x_0)} |D(h^i \vee \psi^i - h^i)|^2\, dx\end{aligned}$$

for $i = 1, \ldots, N$, which implies

$$\int_{B_\rho(x_0)} |D(h \vee \psi - h)|^2\, dx \leq \int_{B_\rho(x_0)} |D\psi - (D\psi)_{x_0,\rho}|^2\, dx \leq c\rho^{n+2\alpha}. \tag{2.15}$$

Combining (2.13), (2.14), and (2.15), we have shown the asserted almost minimality of u. If we only know that the φ^i's are in $C^1(U)$, with a modulus of continuity given by

$$|D\psi(x_0) - D\psi(x)| \leq \mu(|x_0 - x|)\,,$$

the same argument can be applied to show the almost minimality for a function ω given by $\omega(s) = \mu^2(s)$.

Example 3 (almost minimizers of the Dirichlet integral; optimality). As a more general result, we have that every function $u : U \to \mathbb{R}^N$ of class $C^{1,\alpha}$ is an almost minimizer of the Dirichlet integral with $\omega(\rho) = c\rho^{2\alpha}$ for some constant $c > 0$. The proof is a simplified version of the arguments in Example 2, consisting of establishing (2.13) and the inequality

$$\int_{B_\rho(x_0)} |D(u-h)|^2\, dx \leq \int_{B_\rho(x_0)} |Du - (Du)_{x_0,\rho}|^2\, dx \leq c\alpha_n \rho^{n+2\alpha}, \tag{2.16}$$

which is proved exactly like (2.15).

Note in particular that this example shows that our regularity theorem is optimal in the case of Hölder-continuous moduli of continuity. We can in fact show the same for an arbitrary ω satisfying conditions (ω0)-(ω3).

We begin by noting for an arbitrary $u \in C^1(B_\rho(x_0), \mathbb{R}^N)$ that we can combine (2.13) and (2.16) to see

$$\int_{B_\rho(x_0)} |Du|^2\,dx \le \int_{B_\rho(x_0)} |D(u+\varphi)|^2\,dx + \int_{B_\rho(x_0)} |Du - (Du)_{x_0,\rho}|^2\,dx. \tag{2.17}$$

In order to construct our example, we first consider $v : \mathbb{R} \to \mathbb{R}$ given by

$$v(s) = \int_0^s \sqrt{\Omega}(|t|)\,dt\,.$$

We calculate

$$\begin{aligned} \frac{1}{2}\int_{-\rho}^{\rho} |v'(s) - v'_{0,\rho}|^2\,ds &= \int_0^\rho \left|\sqrt{\Omega}(s) - \fint_0^\rho \sqrt{\Omega}(r)\,dr\right|^2 ds \\ &= \int_0^\rho [\sqrt{\Omega}(s)]^2\,ds + \frac{1}{\rho}\left(\int_0^\rho \sqrt{\Omega}(s)\,ds\right)^2 - \frac{2}{\rho}\left(\int_0^\rho \sqrt{\Omega}(s)\,ds\right)^2 \\ &= \int_0^\rho \Omega(s)\,ds - \frac{1}{\rho}\left(\int_0^\rho \sqrt{\Omega}(s)\,ds\right)^2. \end{aligned} \tag{2.18}$$

Since $\sqrt{\omega}(r) = r(\sqrt{\Omega})'(r)$, ($\omega$0) can be expressed as $r(\sqrt{\Omega})'(r) \le s(\sqrt{\Omega})'(s)$ for $r \le s$. Using this in (2.18), we see

$$\begin{aligned} \frac{1}{2}\frac{d}{d\rho}\int_{-\rho}^{\rho} |v'(s) - v'_{0,\rho}|^2\,ds &= \Omega(\rho) + \left(\fint_0^\rho \sqrt{\Omega}(s)\,ds\right)^2 - 2\sqrt{\Omega}(\rho)\fint_0^\rho \sqrt{\Omega}(s)\,ds \\ &= \left(\sqrt{\Omega}(\rho) - \fint_0^\rho \sqrt{\Omega}(s)\,ds\right)^2 \\ &= \left(\fint_0^\rho \left(\int_s^\rho (\sqrt{\Omega})'(t)\,dt\right)\,ds\right)^2 \\ &\le \left(\fint_0^\rho \left(\int_s^\rho \frac{\rho}{t}(\sqrt{\Omega})'(\rho)\,dt\right)\,ds\right)^2 \\ &= \left(\int_0^\rho (\log\rho - \log s)\,ds\right)^2 [(\sqrt{\Omega})'(\rho)]^2 \\ &= \rho^2[(\sqrt{\Omega})'(\rho)]^2 \\ &= \omega(\rho). \end{aligned} \tag{2.19}$$

Integrating this expression, we see

$$\frac{1}{2}\int_{-\rho}^{\rho} |v'(s) - v'_{0,\rho}|^2\,ds \le \int_0^\rho \omega(s)\,ds \le \rho\omega(\rho). \tag{2.20}$$

Consider now a real-valued function u defined on B, the unit ball in $\mathbb{R}^n$, given by

$$u(x) = \int_0^{x_1} \sqrt{\Omega}(|t|)\,dt.$$

In view of ($\omega 3$) we see that $u \in C^1(B)$, and the modulus of continuity of Du is given by $\sqrt{\Omega}$. We consider an arbitrary ball $B_\rho(x_0) \in B$; due to the symmetry of u with respect to x^1, it suffices to consider x_0 with $x_0^1 \geq 0$. We first consider the case that $x_0^1 < 2\rho$. We have, using (2.20) and (2.2),

$$
\begin{aligned}
(2.21) \qquad \int_{B_\rho(x_0)} |Du - (Du)_{x_0,\rho}|^2 \, dx &\leq \alpha_{n-1}\rho^{n-1} \int_{x_0^1-\rho}^{x_0^1+\rho} |v'(s) - v'_{x_0^1,\rho}|^2 \, ds \\
&\leq \alpha_{n-1}\rho^{n-1} \int_{x_0^1-\rho}^{x_0^1+\rho} |v'(s) - v'_{0,x_0^1+\rho}|^2 \, ds \\
&\leq \alpha_{n-1}\rho^{n-1} \int_{-x_0^1-\rho}^{x_0^1+\rho} |v'(s) - v'_{0,x_0^1+\rho}|^2 \, ds \\
&\leq 2\alpha_{n-1}\rho^{n-1}(x_0^1+\rho)\omega(x_0^1+\rho) \\
&\leq 2 \cdot 3^{1+2\alpha}\alpha_{n-1}\rho^n\omega(\rho) \, .
\end{aligned}
$$

For $x_0^1 \geq 2\rho$ we begin by noting that $\sqrt{\Omega}$ is monotone nondecreasing on the interval $(x_0^1 - \rho, x_0^1 + \rho)$. Keeping this in mind, and using ($\omega 1$) twice, we have

$$
\begin{aligned}
\int_{x_0^1-\rho}^{x_0^1+\rho} |v'(s) - v'_{x_0^1,\rho}|^2 \, ds &= \int_{x_0^1-\rho}^{x_0^1+\rho} \left| \sqrt{\Omega}(s) - (\sqrt{\Omega})_{x_0^1,\rho} \right|^2 \, ds \\
&\leq \int_{x_0^1-\rho}^{x_0^1+\rho} \left| \sqrt{\Omega}(s) - \sqrt{\Omega}(x_0^1-\rho) \right|^2 \, ds \\
&\leq \int_{x_0^1-\rho}^{x_0^1+\rho} \left[\int_{x_0^1-\rho}^{s} \frac{\sqrt{\omega}(\sigma)}{\sigma} \, d\sigma \right]^2 \, ds \\
&\leq \frac{\omega(x_0^1-\rho)}{(x_0^1-\rho)^{2\alpha}} \int_{x_0^1-\rho}^{x_0^1+\rho} \left[\int_{x_0^1-\rho}^{s} \frac{d\sigma}{\sigma^{1-\alpha}} \right]^2 \, ds \\
&\leq \frac{\omega(\rho)}{\alpha^2\rho^{2\alpha}} \int_{x_0^1-\rho}^{x_0^1+\rho} [s^\alpha - (x_0^1-\rho)^\alpha]^2 \, ds \\
&\leq \frac{2\rho\omega(\rho)}{\alpha^2\rho^{2\alpha}} [(x_0^1+\rho)^\alpha - (x_0^1-\rho)^\alpha]^2 \\
&\leq 2(3^\alpha - 1)^2\alpha^{-2}\rho\omega(\rho) \, .
\end{aligned}
$$

Hence we have

$$
\begin{aligned}
(2.22) \qquad \int_{B_\rho(x_0)} |Du - (Du)_{x_0,\rho}|^2 \, dx &\leq \alpha_{n-1}\rho^{n-1} \int_{x_0^1-\rho}^{x_0^1+\rho} |v'(s) - v'_{x_0^1,\rho}|^2 \, ds \\
&\leq 2(3^\alpha - 1)^2\alpha^{-2}\alpha_{n-1}\rho^n\omega(\rho) \, .
\end{aligned}
$$

In view of (2.17), the estimates (2.21) and (2.22) show that u is an ω-almost minimizer for the Dirichlet integral on the unit ball B.

Example 4 (volume-constrained minimizers). For a fixed $v_0 \in H^{1,2}(U, \mathbb{R}^N)$ we define $\mathcal{H}_{v_0}$ to be the set of functions v in $H^{1,2}(U, \mathbb{R}^N)$ such that $v = v_0$ on ∂U and $\int_U v \, dx = \int_U v_0 \, dx$. We then consider $u \in \mathcal{H}_{v_0}$ such that

$$
(2.23) \qquad \int_U |Du|^2 \, dx \leq \int_U |Dv|^2 \, dx
$$

for all $v \in \mathcal{H}_{v_0}$; that is, the function u minimizes the Dirichlet integral amongst all functions satisfying a given (vector-valued, signed) *volume constraint*. We will show here that u is an almost minimizer for the Dirichlet integral, for a function $\omega(r) = Cr$ for a suitable constant C. This example was also given by Anzellotti [An, Example 3.2]. In the current situation, due to our more general definition of an almost minimizer (see the comments in the introduction) the calculations are somewhat easier; in particular, in contrast to the result of Anzellotti, the constrained minimizer is an almost minimizer for the same functional. Having said that, we should also state that our calculations are similar to those in [An].

We wish to show for all $x_0 \in U$

$$\int_{B_\rho(x_0)} |Du|^2\,dx \le \int_{B_\rho(x_0)} |D(u+\varphi)|^2 dx + C\rho \int_{B_\rho(x_0)} (1 + |Du|^2 + |D\varphi|^2)\,dx \tag{2.24}$$

for all test functions $\varphi \in H_0^{1,2}(B_\rho(x_0), \mathbb{R}^N)$, for all ρ with $B_\rho(x_0) \subset\subset U$. Define $R_0 = \sup_{x\in U}\{\sup\{r \,|\, B_r(x) \subset\subset U\}\}$, and set $\rho_0 = \rho_0(x_0) = \min\{R_0/4, \operatorname{dist}(x_0, \partial U), 1\}$. Obviously it suffices to establish (2.24) for all ρ with $0 < \rho \le \rho_0$. Let ψ be a fixed function in $H_0^{1,2}(B_{R_0/4}, \mathbb{R}^N)$ with $\int_{B_{R_0/4}} \psi^i \ne 0, i = 1, \dots, N$. We fix $y_0 \in U$ such that $B' = B_{R_0/4}(y_0) \subset U$ and $B' \cap B_\rho(x_0) = \emptyset$. Define $\eta \in H_0^{1,2}(B', \mathbb{R}^N)$ by $\eta(x) = \psi(x - y_0)$.

For a given test function $\varphi \in H_0^{1,2}(B_\rho(x_0), \mathbb{R}^N)$, for $i = 1, \dots, N$ we define $t_i \in \mathbb{R}$ by

$$t_i = \frac{-\int_{B_\rho(x_0)} \varphi^i\,dx}{\int_{B'} \eta^i\,dx} = \frac{-\int_{B_\rho(x_0)} \varphi^i\,dx}{\int_{B_{R_0/4}} \psi^i\,dx}\,. \tag{2.25}$$

Poincaré's inequality yields the estimate

$$|t_i| \le c_3 \rho^{\frac{n}{2}+1} \left(\int_{B_\rho(x_0)} |D\varphi^i|^2\,dx \right)^{1/2} \tag{2.26}$$

for a constant c_3 depending only on n, U, and the fixed function ψ.

We next define a function w via

$$w^i(x) = \begin{cases} u^i(x) + \varphi^i(x), & x \in B_\rho(x_0), \\ u^i(x) + t_i \eta^i(x), & x \in B', \\ u^i(x), & x \in U \setminus (B_\rho(x_0) \cup B') \end{cases}$$

for $i = 1, \dots, N$. We see immediately that $w \in H^{1,2}(U, \mathbb{R}^N)$ and that $w\big|_{\partial U} = u\big|_{\partial U} = v_0\big|_{\partial U}$. From (2.25) we also have that $\int_U w\,dx = \int_U u\,dx$, meaning that $w \in \mathcal{H}_{v_0}$. We thus have from (2.23)

$$\int_{B_\rho(x_0)} |Du|^2\,dx \le \int_{B_\rho(x_0)} |D(u+\varphi)|^2\,dx + \int_{B'} |Du + tD\eta|^2\,dx - \int_{B'} |Du|^2\,dx\,, \tag{2.27}$$

where $tD\eta$ denotes $\{t_i D_\alpha \eta^i\}_{i=1,\dots,N}^{\alpha=1,\dots,n}$. We then estimate

$$\int_{B'} |Du + tD\eta|^2\,dx - \int_{B'} |Du|^2\,dx \le 2\Big|\sum_{i=1}^N \int_{B'} Du^i \cdot D\eta^i\,dx\,\Big| + \sum_{i=1}^N t_i^2 \int_{B'} |D\eta^i|^2\,dx\,.$$

Using (2.26), we see that the second term on the right can be bounded above by $c_4\rho^{n+2}\int_{B_\rho(x_0)} |D\varphi|^2\,dx$ for c_4 depending only on n, U, and ψ. We further have, after applying the Cauchy–Schwarz and then Young inequalities, and taking into account (2.23),

$$\begin{aligned}
2\left|\sum_{i=1}^N t_i \int_{B'} Du^i \cdot D\eta^i\,dx\right| &\le 2\left(\int_U |Du|^2\,dx\right)^{1/2}\left(\sum_{i=1}^N t_i^2 \int_{B'} |D\eta^i|^2\,dx\right)^{1/2}\\
&\le 2\left(\int_U |Dv_0|^2\,dx\right)^{1/2}\left(c_4\rho^{n+2}\int_{B_\rho(x_0)} |D\varphi|^2\,dx\right)^{1/2}\\
&\le c_5\left(\rho^{n+1} + \rho\int_{B_\rho(x_0)} |D\varphi|^2\,dx\right)
\end{aligned}$$

for $c_5 = c_4 + \int_U |Dv_0|^2\,dx$.

Combining these estimates in (2.27), we have (noting that $c_5 \ge c_4$, $\rho \le 1$)

$$\begin{aligned}
\int_{B_\rho(x_0)} |Du|^2\,dx &\le \int_{B_\rho(x_0)} |D(u+\varphi)|^2\,dx + c_5\rho^{n+1} + (c_5\rho + c_4\rho^{n+2})\int_{B_\rho(x_0)} |D\varphi|^2\,dx\\
&\le \int_{B_\rho(x_0)} |D(u+\varphi)|^2\,dx + 2c_5\rho\left(1 + \int_{B_\rho(x_0)} |D\varphi|^2\,dx\right),
\end{aligned}$$

which is the desired estimate.

We also note (again, cf. [An, section 3]) that the same arguments hold for functionals of the form $\int_U A^{\alpha\beta}(x) D_\alpha u D_\beta u\,dx$, under suitable assumptions on the functions $\{A^{\alpha\beta}\}$.

Finally, it should be mentioned here that comparable examples exist in the setting of geometric measure theory; see, e.g., [Al], [Ta], and [DS].

3. The Caccioppoli inequality. We begin by stating an elementary technical lemma from Fusco–Hutchinson, [FH, Lemma 3.2] (cf. [G1, Chapter V, Lemma 3.1]); for completeness we include the result here.

LEMMA 3.1. *Let h be nonnegative and bounded on $[\rho/2, \rho]$, and satisfy*

$$h(t) \le \theta h(s) + A(s-t)^{-2} + B$$

for positive constants A, B, and θ with $0 < \theta < 1$, for all s and t with $\rho/2 \le s < t < \rho$. Then there exists a constant c depending only on θ such that

$$h(\rho/2) \le c(A\rho^{-2} + B)\,.$$

We now prove a suitable version of the Caccioppoli inequality. The proof is close to that of [Ev, Lemma 5.1] and [GM, Proposition 4.1].

LEMMA 3.2. *Let f satisfy* (H2) *and* (H3), *and ω satisfy* (ω0), (ω1), *and* (ω2). *Let F be the functional on $H^{1,2}(U, \mathbb{R}^N)$ given by $F(u) = \int_U f(Du)\,dx$. Then there*

exist positive constants $\rho_1 = \rho_1(\lambda, \omega(\cdot))$ and $c_6 = c_6(\lambda, L)$ (without loss of generality we take $c_6 \geq 1$) such that for every $B_\rho(x_0) \subset\subset U$ with $\rho \leq \rho_1$, $p_0 \in \mathrm{Hom}(\mathbb{R}^n, \mathbb{R}^N)$ and every $u \in H^{1,2}(B_\rho(x_0), \mathbb{R}^N)$ which is (F, ω)-minimizing at x_0 there holds

$$\begin{aligned}(3.1)\qquad &\int_{B_{\rho/2}(x_0)} |Du - p_0|^2\, dx \\ &\leq c_6 \left[\rho^{-2} \int_{B_\rho(x_0)} |u - p_0(x - x_0)|^2\, dx + \alpha_n \omega(\rho) \rho^n (1 + |p_0|^2)\right].\end{aligned}$$

Proof. For $\frac{\rho}{2} \leq t < s \leq \rho$ choose $\eta \in C_0^\infty(B_\rho(x_0), [0,1])$, $\eta \equiv 1$ on $B_t(x_0)$, $\eta \equiv 0$ outside $B_s(x_0)$, and $|\nabla \eta| \leq 2/(s-t)$. We set

$$\begin{aligned}\varphi &:= \eta(u - p_0(x - x_0)), \\ \psi &:= (1 - \eta)(u - p_0(x - x_0)).\end{aligned}$$

Then

$$(3.2)\qquad D\varphi + D\psi = Du - p_0$$

and, with $v(x) := u(x) - p_0(x - x_0)$,

$$(3.3)\qquad |D\varphi|^2 \leq 2|Du - p_0|^2 + \frac{8}{(s-t)^2}|v|^2,$$

$$(3.4)\qquad |D\psi|^2 \leq 2|Du - p_0|^2 + \frac{8}{(s-t)^2}|v|^2.$$

From (H2) and (3.2) we have

$$(3.5)\qquad \lambda \int_{B_s(x_0)} |D\varphi|^2\, dx \leq \int_{B_s(x_0)} [f(p_0 + D\varphi) - f(p_0)]\, dx = I + II + III\,,$$

where

$$\begin{aligned}I &= \int_{B_s(x_0)} [f(Du - D\psi) - f(Du)]\, dx\,, \\ II &= \int_{B_s(x_0)} [f(Du) - f(Du - D\varphi)]\, dx, \qquad \text{and} \\ III &= \int_{B_s(x_0)} [f(p_0 + D\psi) - f(p_0)]\, dx\,.\end{aligned}$$

The (F, ω)-minimality and (3.3), along with $(\omega 2)$, imply

$$\begin{aligned}(3.6)\qquad II &\leq \omega(s) \int_{B_s(x_0)} \left(1 + |Du|^2 + |D\varphi|^2\right) dx \\ &\leq \omega(s) \int_{B_s(x_0)} \left(1 + 2|p_0|^2 + 4|Du - p_0|^2 + \frac{8}{(s-t)^2}|v|^2\right) dx \\ &\leq \frac{\lambda}{2} \int_{B_s(x_0)} |Du - p_0|^2\, dx + \frac{8}{(s-t)^2} \int_{B_s(x_0)} |v|^2\, dx \\ &\quad + 2\alpha_n \omega(\rho) \rho^n (1 + |p_0|^2),\end{aligned}$$

as long as ρ is sufficiently small that $8\omega(\rho) \le \lambda$; by $(\omega 0)$ and $(\omega 1)$ we can choose $\rho_1 > 0$ such that this holds for all $\rho \in (0, \rho_1]$. For the other terms we have (via (2.8) and (3.2), as well as (3.4))

$$
\begin{aligned}
I + III &\le L \int_{B_s(x_0)} \Big(|Du - p_0| + |D\psi|\Big)|D\psi|\, dx \\
&= L \int_{B_s(x_0)\setminus B_t(x_0)} \Big(|Du - p_0| + |D\psi|\Big)|D\psi|\, dx \\
&\le \frac{L}{2} \int_{B_s(x_0)\setminus B_t(x_0)} |Du - p_0|^2\, dx + \frac{3L}{2} \int_{B_s(x_0)\setminus B_t(x_0)} |D\psi|^2\, dx \\
&\le \frac{7L}{2} \int_{B_s(x_0)\setminus B_t(x_0)} |Du - p_0|^2\, dx + \frac{12L}{(s-t)^2} \int_{B_s(x_0)} |v|^2\, dx.
\end{aligned} \tag{3.7}
$$

Combining (3.6) and (3.7) in (3.5) and noting $D\varphi = Du - p_0$ on $B_t(x_0)$ we see

$$
\begin{aligned}
\frac{\lambda}{2} \int_{B_t(x_0)} |Du - p_0|^2\, dx &\le \frac{7L + \lambda}{2} \int_{B_s(x_0)\setminus B_t(x_0)} |Du - p_0|^2\, dx \\
&+ \frac{12L + 8}{(s-t)^2} \int_{B_s(x_0)} |v|^2\, dx + 2\alpha_n \omega(\rho) \rho^n (1 + |p_0|^2).
\end{aligned} \tag{3.8}
$$

Thus we have

$$
\begin{aligned}
\int_{B_t(x_0)} |Du - p_0|^2\, dx &\le \frac{7L + \lambda}{7L + 2\lambda} \int_{B_s(x_0)} |Du - p_0|^2\, dx \\
&+ \frac{24L + 16}{7L(s-t)^2} \int_{B_s(x_0)} |v|^2\, dx + \frac{4}{7L} \alpha_n \omega(\rho) \rho^n (1 + |p_0|^2).
\end{aligned} \tag{3.9}
$$

Since $\frac{7L+\lambda}{7L+2\lambda} < 1$ we can apply Lemma 3.1 to conclude (3.1). ∎

4. Approximate A-harmonicity and A-harmonic approximation. The next lemma is a prerequisite for applying the A-harmonic approximation technique.

LEMMA 4.1. *Let ω satisfy $(\omega 2)$, and f satisfy* (H2), (H3), *and* (H4). *Let F be the functional on $H^{1,2}(U, \mathbb{R}^N)$ given by $F(u) = \int_U f(Du)\, dx$. Then there exists $c_7 = c_7(n, L)$ such that for every $u \in H^{1,2}(U, \mathbb{R}^N)$ that is (F, ω)-minimizing at x_0, every ball $B_\rho(x_0) \subset\subset U$, and every $p_0 \in \mathrm{Hom}(\mathbb{R}^n, \mathbb{R}^N)$ we have*

$$
\begin{aligned}
&\left| \rho^{-n} \int_{B_\rho(x_0)} D^2 f(p_0)(Du - p_0, D\varphi)\, dx \right| \\
&\le c_7 \Big[\omega^{1/2}(\rho)(1 + \Phi + |p_0|^2) + \nu^{1/2}(\Phi)\Phi^{1/2} \Big] \sup_{B_\rho(x_0)} |D\varphi|
\end{aligned} \tag{4.1}
$$

for all $\varphi \in C_0^1(B_\rho(x_0), \mathbb{R}^N)$. Here we write

$$
\Phi = \Phi(x_0, \rho, p_0) := \fint_{B_\rho(x_0)} |Du - p_0|^2\, dx. \tag{4.2}
$$

Proof. Without loss of generality we take $x_0 = 0$. We first note

$$
\begin{aligned}
\int_{B_\rho} Df(Du) \cdot D\varphi\, dx &= \int_{B_\rho} Df(Du) \cdot D\varphi\, dx - \int_{B_\rho} Df(p_0) \cdot D\varphi\, dx \\
&= \int_{B_\rho} \int_0^1 D^2 f(p_0 + \tau(Du - p_0))\, (Du - p_0, D\varphi) d\tau dx.
\end{aligned} \tag{4.3}
$$

Initially we assume $|D\varphi| \le 1$ on B_ρ. For positive s we have from the (F,ω)-minimality of u

$$
\begin{aligned}
(4.4)\quad & \int_{B_\rho} D^2 f(p_0)(Du - p_0, D\varphi)\,dx \\
&\ge \frac{1}{s}\left[\int_{B_\rho} (f(Du) - f(Du + sD\varphi))\,dx - \omega(\rho)\int_{B_\rho}\left(1 + |Du|^2 + s^2|D\varphi|^2\right)dx\right] \\
&\qquad + \int_{B_\rho} D^2 f(p_0)(Du - p_0, D\varphi)\,dx \\
&\ge \frac{1}{s}\left[-\int_{B_\rho}\int_0^s \frac{d}{dt} f(Du + tD\varphi)\,dt\,dx + s\int_{B_\rho} D^2 f(p_0)(Du - p_0, D\varphi)\,dx\right. \\
&\qquad \left. - \omega(\rho)\int_{B_\rho}\left(1 + s^2 + |Du|^2\right)dx\right] \qquad \text{since } |D\varphi| \le 1 \\
&= \frac{1}{s}\left[\int_{B_\rho}\int_0^s \Big(Df(Du) - Df(Du + tD\varphi)\Big)\cdot D\varphi\,dt\,dx\right. \\
&\qquad + s\int_{B_\rho}\int_0^1 \Big(D^2 f(p_0) - D^2 f(p_0 + \tau(Du - p_0))\Big)\,d\tau(Du - p_0, D\varphi)\,dx \\
&\qquad \left. - \omega(\rho)\int_{B_\rho}\left(1 + s^2 + |Du|^2\right)dx\right] \qquad \text{via (4.3)} \\
&\ge -\frac{1}{s}\left[\frac{L}{2}s^2\alpha_n\rho^n + s\sqrt{2L}\int_{B_\rho}\nu^{1/2}(|Du - p_0|^2)|Du - p_0|\,dx\right. \\
&\qquad \left. + \omega(\rho)\int_{B_\rho}\left(1 + s^2 + |Du|^2\right)dx\right] \qquad \text{via (2.8), (2.10), (H3)} \\
&\ge -\frac{L}{2}s\alpha_n\rho^n\,dx - \sqrt{2L}\,\alpha_n\rho^n\nu^{1/2}\left(\fint_{B_\rho}|Du - p_0|^2\,dx\right)\left(\fint_{B_\rho}|Du - p_0|^2\,dx\right)^{1/2} \\
&\qquad - \frac{\omega(\rho)}{s}\int_{B_\rho}\left(1 + s^2 + 2|Du - p_0|^2 + 2|p_0|^2\right)dx \\
&\ge -\alpha_n\rho^n\left[\frac{L}{2}s + \sqrt{2L}\,\nu^{1/2}(\Phi)\Phi^{1/2} + \frac{2\omega(\rho)}{s}(1 + s^2 + \Phi + |p_0|^2)\right];
\end{aligned}
$$

we have used the Jensen and Hölder inequalities to obtain the second to last inequality. Completely analogously we see

$$
\begin{aligned}
(4.5)\quad & \int_{B_\rho} D^2 f(p_0)(Du - p_0, D\varphi)\,dx \\
&\qquad \le \alpha_n\rho^n\left[\frac{L}{2}s + \sqrt{2L}\,\nu^{1/2}(\Phi)\Phi^{1/2} + \frac{2\omega(\rho)}{s}(1 + s^2 + \Phi + |p_0|^2)\right].
\end{aligned}
$$

By choosing $s := \omega^{1/2}(\rho)$ and using $(\omega 2)$ we have the desired conclusion for φ such that $|D\varphi| \le 1$ with $c_7 = \alpha_n(4 + L)$. By a simple scaling argument this yields the result for general φ. □

We close this section by giving a result which is central to our technique, the A-harmonic approximation lemma. The lemma was first proven in [DS, Lemma 3.3];

cf. [S2, section 1.6] for the case $A = id$ (i.e., the harmonic approximation lemma); for completeness, we quote it here.

LEMMA 4.2. *Consider fixed positive λ and L, and n, $N \in \mathbb{N}$ with $n \geq 2$. Then for any given $\varepsilon > 0$ there exists $\delta = \delta(n, N, \lambda, L, \varepsilon) \in (0, 1]$ with the following property: if $A \in \mathrm{Bil}(\mathrm{Hom}(\mathbb{R}^n, \mathbb{R}^N))$ is rank-one elliptic with ellipticity constant $\lambda > 0$ and upper bound L, then for any $u \in H^{1,2}(B_\rho(x_0), \mathbb{R}^N)$ (for some $\rho > 0$, $x_0 \in \mathbb{R}^n$) satisfying*

$$\rho^{-n} \int_{B_\rho(x_0)} |Du|^2 \, dx \leq 1, \qquad \text{and}$$

$$\left| \rho^{-n} \int_{B_\rho(x_0)} A(Du, D\varphi) \, dx \right| \leq \delta(n, N, \lambda, L, \varepsilon) \sup |D\varphi|$$

for all $\varphi \in C_0^1(B_\rho(x_0), \mathbb{R}^N)$, there exists an A-harmonic function $h \in H^{1,2}(B_\rho(x_0), \mathbb{R}^N)$ such that

$$\rho^{-n} \int_{B_\rho} |Dh|^2 \, dx \leq 1 \qquad \text{and} \quad \rho^{-n-2} \int_{B_\rho} |h - u|^2 \, dx \leq \varepsilon.$$

Here h is called *A-harmonic* if

$$\int_{B_\rho} A(Dh, D\varphi) \, dx = 0$$

for all $\varphi \in C_0^\infty(B_\rho(x_0), \mathbb{R}^N)$.

5. Proof of the main theorem. To prove the result we follow the general lines of [DG, section 3]. We first establish appropriate smallness conditions sufficient to deduce growth estimates on Φ.

PROPOSITION 5.1. *Consider u satisfying the conditions of Theorem 2.2, and β fixed, $\alpha < \beta < 1$. We write $\Phi(x_0, r)$ for $\Phi(x_0, r, (Du)_{x_0,r})$. Then we can find positive constants c_8, c_9, and δ, and $\theta \in (0, 1)$ (with c_8 depending only on n, N, λ, and L, and with c_9, θ, and δ depending only on these quantities as well as β) such that the smallness conditions $\rho \leq \rho_1$,*

$$\nu(\Phi(x_0, \rho)) + \Phi(x_0, \rho) \leq \delta^2/2\,, \tag{5.1}$$

and

$$c_8 \omega(\rho)(1 + |(Du)_{x_0,\rho}|^4) \leq \delta^2 \tag{5.2}$$

together imply the growth condition

$$\Phi(x_0, \theta\rho) \leq \theta^{2\beta} \Phi(x_0, \rho) + c_9 \omega(\rho)(1 + |(Du)_{x_0,\rho}|^4). \tag{5.3}$$

Here ρ_1 depending on λ and $\omega(\cdot)$ is given in Lemma 3.2.

Proof. From Lemma 4.1 we have (with $c_{10} := 1 + \sqrt{2} c_7$)

$$\begin{aligned} \left| \rho^{-n} \int_{B_\rho(x_0)} D^2 f(p_0)(Du - p_0, D\varphi) \, dx \right| &\leq c_{10} \Big[\Phi(x_0, \rho, p_0) \\ &+ \nu^{1/2}(\Phi(x_0, \rho, p_0)) \Phi^{1/2}(x_0, \rho, p_0) + (\omega(\rho)/2)^{1/2}(1 + |p_0|^2) \Big] \sup_{B_\rho(x_0)} |D\varphi|. \end{aligned} \tag{5.4}$$

We set

$$w = \frac{u - p_0(x - x_0)}{2c_{10}\sqrt{\Phi(x_0, \rho, p_0) + \delta^{-2}\omega(\rho)(1 + |p_0|^2)^2}} \tag{5.5}$$

and deduce from (5.4) that for all $\varphi \in C_c^\infty(B_\rho(x_0), \mathbb{R}^n)$ there holds

$$\begin{aligned} &\left|\rho^{-n}\int_{B_\rho(x_0)} D^2 f(p_0)(Dw, D\varphi)\,dx\right| \\ &\le \tfrac{1}{2}\Big[\Phi^{1/2}(x_0, \rho, p_0) + \nu^{1/2}(\Phi(x_0, \rho, p_0)) + \delta/\sqrt{2}\Big] \sup_{B_\rho(x_0)} |D\varphi| \\ &\le \Big[\nu(\Phi(x_0, \rho, p_0)) + \Phi(x_0, \rho, p_0) + \delta^2/2\Big]^{1/2} \sup_{B_\rho(x_0)} |D\varphi| \end{aligned} \tag{5.6}$$

and (since $c_{10} \ge \max\{\alpha_n, 1\}$),

$$\rho^{-n}\int_{B_\rho(x_0)} |Dw|^2\,dx \le \frac{\alpha_n}{4c_{10}^2} \le 1. \tag{5.7}$$

We further set

$$A(\xi, \eta) := D^2 f(p_0)(\xi, \eta). \tag{5.8}$$

From (2.7) we see that the bilinear form A satisfies the conditions of Lemma 4.2. For positive ε to be determined later, we denote by $\delta = \delta(n, N, \lambda, L, \varepsilon) \in (0, 1]$ the corresponding constant from Lemma 4.2; via this lemma the *smallness condition*

$$\nu(\Phi(x_0, \rho, p_0)) + \Phi(x_0, \rho, p_0) \le \delta^2/2 \tag{5.9}$$

guarantees the existence of an A-harmonic $h \in H^{1,2}(B_\rho(x_0), \mathbb{R}^N)$ satisfying

$$\rho^{-n}\int_{B_\rho(x_0)} |Dh|^2\,dx \le 1 \qquad \text{and} \tag{5.10}$$

$$\rho^{-n-2}\int_{B_\rho(x_0)} |w - h|^2\,dx \le \varepsilon. \tag{5.11}$$

We also note that h satisfies the estimate

$$\rho^{-2}\sup_{B_{\rho/2}(x_0)} |Dh|^2 + \sup_{B_{\rho/2}(x_0)} |D^2 h|^2 \le c_{11}\rho^{-n-2}\int_{B_\rho(x_0)} |Dh|^2 dx \le \frac{c_{11}}{\rho^2}\,, \tag{5.12}$$

with $c_{11} = c_{11}(n, N, \lambda, L)$ (without loss of generality we take $c_{11} \ge 1$). For elliptic A the first inequality follows from a standard argument due to Campanato (see [Ca, Teorema 9.2]) combined with the Sobolev and Poincaré inequalities; the same arguments are valid in the current setting because the Legendre–Hadamard condition is satisfied; cf. [Ev, p. 236]. The second inequality follows from (5.10). For $\theta \in (0, 1/4]$ we can thus apply Taylor's theorem to h at x_0 to deduce

$$\sup_{x \in B_{2\theta\rho}(x_0)} |h(x) - h(x_0) - Dh(x_0)(x - x_0)|^2 \le \frac{c_{11}}{\rho^2}(2\theta\rho)^4 = 16c_{11}\theta^4\rho^2.$$

Thus we have, using also (5.11),

$$(2\theta\rho)^{-n-2}\int_{B_{2\theta\rho}(x_0)}|w-h(x_0)-Dh(x_0)(x-x_0)|^2dx \tag{5.13}$$
$$\le 2(2\theta\rho)^{-n-2}\left(\int_{B_{2\theta\rho}(x_0)}|w-h|^2dx+\int_{B_{2\theta\rho}(x_0)}|h-h(x_0)-Dh(x_0)(x-x_0)|^2dx\right)$$
$$\le 2(2\theta\rho)^{-n-2}(\rho^{n+2}\varepsilon+16c_{11}\alpha_n(2\theta\rho)^n\theta^4\rho^2)$$
$$= 2^{-n-1}\theta^{-n-2}\varepsilon+8c_{11}\alpha_n\theta^2.$$

We now set $\gamma = 2c_{10}\sqrt{\Phi(x_0,\rho,p_0)+\delta^{-2}\omega(\rho)(1+|p_0|^2)^2}$. Taking advantage of the fact that u and $u-(p_0+\gamma Dh(x_0))(x-x_0)$ have the same mean value on balls centered at x_0 we have

$$(2\theta\rho)^{-n-2}\int_{B_{2\theta\rho}(x_0)}|u-u_{x_0,2\theta\rho}-(p_0+\gamma Dh(x_0))(x-x_0)|^2dx \tag{5.14}$$
$$\le (2\theta\rho)^{-n-2}\int_{B_{2\theta\rho}(x_0)}|u-p_0(x-x_0)-\gamma(h(x_0)+Dh(x_0)(x-x_0))|^2dx$$
$$= \gamma^2(2\theta\rho)^{-n-2}\int_{B_{2\theta\rho}(x_0)}|w-h(x_0)-Dh(x_0)(x-x_0)|^2dx$$
$$\le 4c_{10}^2\left(2^{-n-1}\theta^{-n-2}\varepsilon+8c_{11}\alpha_n\theta^2\right)\left(\Phi(x_0,\rho,p_0)+\delta^{-2}\omega(\rho)(1+|p_0|^2)^2\right)$$
$$\le c_{12}\left(\theta^{-n-2}\varepsilon+\theta^2\right)\left(\Phi(x_0,\rho,p_0)+\delta^{-2}\omega(\rho)(1+|p_0|^2)^2\right),$$

where we have used (5.13) in the second to last line; here we have set $c_{12}=(2^{1-n}+32\alpha_nc_{11})c_{10}^2+1$, which depends only on n, N, λ, and L. We now fix $p_0=(Du)_{x_0,\rho}$. With $P=(Du)_{x_0,\rho}+\gamma Dh(x_0)$ we deduce from (5.14), assuming $\rho\le\rho_1$,

$$\Phi(x_0,\theta\rho)=\alpha_n^{-1}(\theta\rho)^{-n}\int_{B_{\theta\rho}(x_0)}|Du-(Du)_{x_0,\theta\rho}|^2\,dx$$
$$\le \alpha_n^{-1}(\theta\rho)^{-n}\int_{B_{\theta\rho}(x_0)}|Du-P|^2\,dx$$
$$\le 2^nc_6\alpha_n^{-1}(2\theta\rho)^{-n-2}\int_{B_{2\theta\rho}(x_0)}|u-u_{x_0,2\theta\rho}-P(x-x_0)|^2\,dx$$
$$+2^nc_6\omega(2\theta\rho)(1+|P|^2)$$
$$\le 2^nc_6c_{12}\alpha_n^{-1}\left(\theta^{-n-2}\varepsilon+\theta^2\right)\left(\Phi(x_0,\rho)+\delta^{-2}\omega(\rho)(1+|(Du)_{x_0,\rho}|^2)^2\right)$$
$$+2^nc_6\omega(\rho)(1+|P|^2)\,; \tag{5.15}$$

here the second to last inequality follows from Lemma 3.2, the last from (5.14). Under the additional smallness condition

$$2c_{11}\gamma^2\le 1 \tag{5.16}$$

we have, using (5.10) and (5.12),

$$1+|P|^2\le 1+2|(Du)_{x_0,\rho}|^2+2\gamma^2|Dh(x_0)|^2 \tag{5.17}$$
$$\le 1+2|(Du)_{x_0,\rho}|^2+2c_{11}\gamma^2\rho^{-n}\int_{B_\rho(x_0)}|Dh|^2\,dx$$
$$\le 1+2|(Du)_{x_0,\rho}|^2+2c_{11}\gamma^2$$
$$\le 2(1+|(Du)_{x_0,\rho}|^2)\,.$$

We now fix θ sufficiently small that

$$2^{n+1}c_6c_{12}\alpha_n^{-1}\theta^2 \le \theta^{2\beta}, \tag{5.18}$$

and then set $\varepsilon := \theta^{n+4}$, which also fixes δ; without loss of generality we assume that δ is sufficiently small that we have $8c_{10}^2c_{11}\delta^2 < 1$. Note that θ, ε, and δ depend on n, N, λ, L, α, and β.

We now set $c_8 = 32c_{10}^2c_{11}$ and $c_9 = 2^{n+2}c_6(\delta^{-2}+1)$. In view of the smallness conditions (5.9), (5.16), and (5.18), inequalities (5.15) and (5.17) then yield the desired result. □

For a given $M > 0$ we can find $\Phi_0(M) > 0$ (dependent also on n, N, λ, L, β, and $\nu(\,\cdot\,)$) sufficiently small that

$$\nu(2\Phi_0(M)) + 2\Phi_0(M) \le \delta^2/2 \qquad \text{and} \tag{5.19}$$

$$\Phi_0(M) \le \frac{1}{4}M^2\theta^n(1-\theta^\beta)^2. \tag{5.20}$$

Given this, we can also find $\rho_0(M) \in (0, \rho_1]$ (dependent also on n, N, λ, L, β, $\nu(\,\cdot\,)$ and $\omega(\,\cdot\,)$) so small that, writing $c_{13}(M)$ for $\frac{c_8+c_9}{\theta^{2\alpha}-\theta^{2\beta}}(1+16M^4)$ (with c_{13} thus depending also on n, N, λ, L, α and β), we have

$$c_{13}(M)\omega(\rho_0(M)) \le \min\{\delta^2, \Phi_0(M)\} \qquad \text{and} \tag{5.21}$$

$$c_{13}(M)\Omega(\rho_0(M)) \le \tfrac{1}{4}M^2\theta^n(1-\theta^\alpha)^2. \tag{5.22}$$

If the quantities $\Phi(x_0,\rho)$ and ρ are sufficiently small for some $B_\rho(x_0)$, the next lemma shows that we can iterate Proposition 5.1.

LEMMA 5.2. *For $M_0 > 0$ and $B_\rho(x_0) \subset\subset U$, suppose that the conditions*

$$\begin{aligned} &\text{(i)} && |(Du)_{x_0,\rho}| \le M_0, \\ &\text{(ii)} && \rho \le \rho_0(M_0), && \textit{and} \\ &\text{(iii)} && \Phi(x_0,\rho) \le \Phi_0(M_0) \end{aligned}$$

are satisfied. Then the smallness conditions (5.1) *and* (5.2) *are fulfilled on $B_{\theta^j\rho}(x_0)$ for all $j \in \mathbb{N}$. Furthermore there exists*

$$\Upsilon_{x_0} := \lim_{j\to\infty}(Du)_{x_0,\theta^j\rho},$$

and there exists c_{14} depending only on n, N, λ, L, α, β, and M_0 such that for all $r < \rho$ there holds

$$\fint_{B_r(x_0)} |Du - \Upsilon_{x_0}|^2\,dx \le c_{14}\Big(\Big(\frac{r}{\rho}\Big)^{2\beta}\Phi(x_0,\rho) + \Omega(r)\Big). \tag{5.23}$$

Proof. In order to show the first part of the lemma we prove two statements by induction. Precisely, for $j \in \mathbb{N}\cup\{0\}$ we shall show

$$\begin{aligned} &\mathbf{(I)}_j && \Phi(x_0,\theta^j\rho) \le \theta^{2\beta j}\Phi(x_0,\rho) + c_{13}(M_0)\omega(\theta^j\rho) && \text{and} \\ &\mathbf{(II)}_j && |(Du)_{x_0,\theta^j\rho}| \le 2M_0\,. \end{aligned}$$

Note first that $(\mathrm{II})_\mathrm{j}$ combined with (5.21) and (iii) yields

$$(\mathbf{I}')_\mathrm{j} \qquad \Phi(x_0,\theta^j\rho)\le 2\Phi_0(M_0)\,.$$

We now proceed to the proof by induction. The case $j=0$ follows immediately from (5.19), (5.21), and the monotonicity of ν and of ω. We assume $(\mathrm{I})_\ell$ and $(\mathrm{II})_\ell$ for $\ell=0,\dots,j-1$. We first calculate, using (5.3), $(\mathrm{II})_\ell$ for $\ell=0,\dots,j-1$ and $(\omega 1)$,

$$\begin{aligned}
\Phi(x_0,\theta^j\rho)&\le \theta^{2\beta j}\Phi(x_0,\rho)+c_9\sum_{\ell=0}^{j-1}\theta^{2\beta\ell}\omega(\theta^{j-\ell-1}\rho)(1+|(Du)_{x_0,\theta^{j-\ell-1}\rho}|^4)\\
&\le \theta^{2\beta j}\Phi(x_0,\rho)+c_9\theta^{-2\alpha}\left(\sum_{\ell=0}^{j-1}\theta^{2(\beta-\alpha)\ell}\right)\omega(\theta^j\rho)(1+16M_0^4)\\
&\le \theta^{2\beta j}\Phi(x_0,\rho)+\frac{c_9(1+16M_0^4)}{\theta^{2\alpha}-\theta^{2\beta}}\omega(\theta^j\rho)\\
&\le \theta^{2\beta j}\Phi(x_0,\rho)+c_{13}(M_0)\omega(\theta^j\rho)\,,
\end{aligned}$$

showing $(\mathrm{I})_\mathrm{j}$. To show $(\mathrm{II})_\mathrm{j}$ we estimate

$$\begin{aligned}
|(Du)_{x_0,\theta^j\rho}|&\le M_0+\sum_{\ell=1}^{j}|(Du)_{x_0,\theta^\ell\rho}-(Du)_{x_0,\theta^{\ell-1}\rho}| && \text{via (iii)}\\
&\le M_0+\sum_{\ell=1}^{j}\left[\fint_{B_{\theta^\ell\rho}(x_0)}|Du-(Du)_{x_0,\theta^{\ell-1}\rho}|^2\,dx\right]^{1/2}\\
&\le M_0+\theta^{-n/2}\sum_{\ell=1}^{j}\left[\fint_{B_{\theta^{\ell-1}\rho}(x_0)}|Du-(Du)_{x_0,\theta^{\ell-1}\rho}|^2\,dx\right]^{1/2}\\
&\le M_0+\theta^{-n/2}\sum_{\ell=0}^{j-1}\sqrt{\theta^{2\beta\ell}\Phi(x_0,\rho)+c_{13}(M_0)\omega(\theta^\ell\rho)} && \text{via (I)}_\ell,\ \ell=0,\dots,j-1\\
&\le M_0+\theta^{-n/2}\left(\frac{\sqrt{\Phi(x_0,\rho)}}{1-\theta^\beta}+\frac{\sqrt{c_{13}(M_0)}}{1-\theta^\alpha}\sqrt{\Omega(\rho)}\right) && \text{via (2.5)}\\
&\le M_0+\theta^{-n/2}\left(\frac{\sqrt{\Phi_0(M_0)}}{1-\theta^\beta}+\frac{\sqrt{c_{13}(M_0)\Omega(\rho_0(M_0))}}{1-\theta^\alpha}\right) && \text{via (iii), (ii)}\\
&\le 2M_0 && \text{via (5.20), (5.22).}
\end{aligned}$$

The conclusion of the lemma then follows from $(\mathrm{I}')_\mathrm{j}$ and $(\mathrm{II})_\mathrm{j}$ after taking into account (5.19) and (5.21).

Analogously we calculate, for $k>j$,

$$\begin{aligned}
|(Du)_{x_0,\theta^j\rho}-(Du)_{x_0,\theta^k\rho}|&\le\sum_{\ell=j+1}^{k}|(Du)_{x_0,\theta^\ell\rho}-(Du)_{x_0,\theta^{\ell-1}\rho}|\\
&\le\theta^{-n/2}\left(\frac{\sqrt{\Phi(x_0,\rho)}}{1-\theta^\beta}\theta^{\beta j}+\frac{\sqrt{c_{13}(M_0)}}{1-\theta^\alpha}\sqrt{\Omega(\theta^j\rho)}\right);
\end{aligned}$$

this shows that $\{(Du)_{x_0,\theta^j\rho}\}$ is a Cauchy sequence. For

$$\Upsilon_{x_0}:=\lim_{j\to\infty}(Du)_{x_0,\theta^j\rho}$$

we thus have, with $c_{15} = \sqrt{2}\theta^{-n/2}\Big(\frac{1+\sqrt{c_{13}(M_0)}}{1-\theta^\alpha}\Big)$ depending only on n, N, λ, L, α, β, and M_0,

$$|(Du)_{x_0,\theta^j\rho} - \Upsilon_{x_0}| \le c_{15}\Big[\theta^{2\beta j}\Phi(x_0,\rho) + \Omega(\theta^j\rho)\Big]^{1/2}$$

for all j. Combining this with $(\mathrm{I})_\mathrm{j}$ and setting $c_{16} = 2(c_{13}(M_0) + c_{15}^2)$ (note that c_{16} has the same dependencies as c_{15}) we have, using also (2.6),

$$\begin{aligned}
&\fint_{B_{\theta^j\rho}(x_0)} |Du - \Upsilon_{x_0}|^2\,dx \le 2\Phi(x_0,\theta^j\rho) + 2|(Du)_{x_0,\theta^j\rho} - \Upsilon_{x_0}|^2\\
&\le 2\theta^{2\beta j}\Phi(x_0,\rho) + 2c_{13}(M_0)\omega(\theta^j\rho) + 2c_{15}^2\Big(\theta^{2\beta j}\Phi(x_0,\rho) + \Omega(\theta^j\rho)\Big)\\
&\le c_{16}\Big(\theta^{2\beta j}\Phi(x_0,\rho) + \Omega(\theta^j\rho)\Big).
\end{aligned}$$

For $0 < r \le \rho$ we can find $j \in \mathbb{N}\cup\{0\}$ with $\theta^{j+1}\rho < r \le \theta^j\rho$. For this j we have

$$\begin{aligned}
&\fint_{B_r(x_0)} |Du - \Upsilon_{x_0}|^2\,dx \le \theta^{-n}\fint_{B_{\theta^j\rho}(x_0)} |Du - \Upsilon_{x_0}|^2\,dx\\
&\le c_{16}\theta^{-n}\Big(\theta^{2j\beta}\Phi(x_0,\rho) + \Omega(\theta^j\rho)\Big)\\
&= c_{16}\theta^{-n}\left(\frac{\theta^{2(j+1)\beta}}{\theta^{2\beta}}\Phi(x_0,\rho) + \Omega\left(\frac{\theta^{j+1}\rho}{\theta}\right)\right)\\
&\le c_{16}\theta^{-n}\Big(\Big(\frac{r}{\rho}\Big)^{2\beta}\theta^{-2\beta}\Phi(x_0,\rho) + \theta^{-2\alpha}\Omega(\theta^{j+1}\rho)\Big)\\
&\le c_{16}\theta^{-n-2\beta}\Big(\Big(\frac{r}{\rho}\Big)^{2\beta}\Phi(x_0,\rho) + \Omega(r)\Big);
\end{aligned} \tag{5.24}$$

here we have used (2.3) to obtain the second to last inequality. This shows (5.23) with $c_{14} = c_{16}\theta^{-n-2\beta}$ (note that c_{14} has the correct dependencies). ∎

We are now in a position to complete the partial-regularity proof.

Proof of Theorem 2.2. We give the proof of (i); the proof of (ii) is completely analogous. We assume that for some $x_0 \in U$ and $M_0 > 0$ we have

$$|(Du)_{x_0,\rho}| < M_0 \quad\text{and}\quad \Phi(x_0,\rho) < \Phi_0(M_0)$$

on $B_\rho(x_0)$, where $B_{2\rho}(x_0) \subset\subset U$ with $0 < \rho \le \rho_0(M_0)$. Such a ρ can always be found for each x_0 belonging neither to Σ_1 nor to Σ_2. Since the functions $z \mapsto (Du)_{z,\rho}$ and $z \mapsto \Phi(z,\rho)$ are continuous there exists a ball $B_\sigma(x_0) \subset\subset U$, such that for all $z \in B_\sigma(x_0)$ we have $B_\rho(z) \subset\subset U$, and further there holds

$$|(Du)_{z,\rho}| < M_0 \quad\text{and}\quad \Phi(z,\rho) < \Phi_0(M_0) \quad \text{for all } z \in B_\sigma(x_0). \tag{5.25}$$

We can thus apply Lemma 5.2 on $B_r(z)$ for any $z \in B_\sigma(x_0)$ and r with $0 < r \le \rho$ to deduce

$$\fint_{B_r(z)} |Du - \Upsilon_z|^2\,dx \le c_{14}\Big(\Big(\frac{r}{\rho}\Big)^{2\beta}\Phi(z,\rho) + \Omega(r)\Big). \tag{5.26}$$

For $z, \widetilde{z} \in B_\sigma(x_0)$ with $r = |z - \widetilde{z}| < 2\sigma$ and $a = (z + \widetilde{z})/2$ we obtain

$$\begin{aligned}
|\Upsilon_z - \Upsilon_{\widetilde{z}}|^2 &= \frac{1}{\alpha_n (r/2)^n} \int_{B_{r/2}(a)} |\Upsilon_z - \Upsilon_{\widetilde{z}}|^2 \, dx \\
&\le \frac{2^n}{\alpha_n r^n} \int_{B_r(z) \cap B_r(\widetilde{z})} |\Upsilon_z - \Upsilon_{\widetilde{z}}|^2 \, dx \\
&\le 2^{n+1} \left[\fint_{B_r(z)} |Du - \Upsilon_z|^2 \, dx + \fint_{B_r(\widetilde{z})} |Du - \Upsilon_{\widetilde{z}}|^2 \, dx \right] \\
&\le 2^{n+1} c_{14} \left[\Big(\frac{r}{\rho}\Big)^{2\beta} \Big(\Phi(z, \rho) + \Phi(\widetilde{z}, \rho) \Big) + 2\Omega(r) \right] \\
&\le 2^{2n+2} c_{14} \left[\Big(\frac{|z - \widetilde{z}|}{\rho} \Big)^{2\beta} \Phi(x_0, 2\rho) + \Omega(|z - \widetilde{z}|) \right].
\end{aligned}$$

Here we have used (5.26) in the third to last inequality, and the fact that $\Phi(z, \rho) + \Phi(\widetilde{z}, \rho) \le 2^{n+1} \Phi(x_0, 2\rho)$ in obtaining the final inequality. Since Υ_z is the Lebesgue-representative of $Du(z)$, we can conclude the desired continuity. □

As noted in section 2 we can weaken the hypotheses of the theorem by omitting (H4). This entails essentially only notational changes in the proof: in (5.1), (5.9), and (5.18) we need to replace $\nu(\,\cdot\,)$ by $\nu(M + 1, \,\cdot\,)$ for $|(Du)_{x_0,\rho}|$ (respectively, $|p_0|$) less than M and check that this is preserved in the iteration. Analogous changes also need to be made in Lemma 4.1. We thus have the following corollary.

COROLLARY 5.3. *The conclusion of Theorem* 2.2 *also follows if we omit the hypothesis* (H4).

REFERENCES

[A] W. K. ALLARD, *On the first variation of a varifold*, Ann. of Math. 2, 95 (1972), pp. 417–491.

[Al] F. J. ALMGREN, *Existence and regularity almost everywhere of solutions to elliptic variational problems with constraints*, Mem. Amer. Math. Soc., 4 (1976), pp. 1–199.

[An] G. ANZELLOTTI, *On the $C^{1,\alpha}$–regularity of ω–minima of quadratic functionals*, Boll. Un. Mat. Ital. C (6), 2 (1983), pp. 195–212.

[AF] E. ACERBI AND N. FUSCO, *Semicontinuity problems in the calculus of variations*, Arch. Rational Mech. Anal., 86 (1984), pp. 125–145.

[Ba] J. M. BALL, *Convexity conditions and existence theorems in nonlinear elasticity*, Arch. Rational Mech. Anal., 63 (1977), pp. 337–403.

[Bo] E. BOMBIERI, *Regularity theory for almost minimal currents*, Arch Rational Mech. Anal., 78 (1982), pp. 99–130.

[Ca] S. CAMPANATO, *Equazioni ellitiche del II^e ordine e spazi $\mathcal{L}^{2,\lambda}$*, Ann. Mat. Pura Appl. (4), 69 (1965), pp. 321–381.

[Da] B. DACOROGNA, *Direct Methods in the Calculus of Variations*, Springer, Berlin, Heidelberg, New York, 1989.

[DeG] E. DE GIORGI, *Un esempio di estremali discontinue per un problema variazionale di tipo ellitico,* Boll. Un. Mat. Ital. (4), 1 (1968), pp. 135–137.

[DG] F. DUZAAR AND J. F. GROTOWSKI, *Partial regularity for nonlinear elliptic systems: The method of A-harmonic approximation*, Manuscripta Math., to appear.

[DS] F. DUZAAR AND K. STEFFEN, *Optimal Interior and Boundary Regularity for Almost Minimizers to Elliptic Integrands*, preprint.

[Ev] L. C. EVANS, *Quasiconvexity and partial regularity in the calculus of variations*, Arch. Rational Mech. Anal., 95 (1986), pp. 227–252.

[Fe] H. FEDERER, *Geometric Measure Theory*, Springer, Berlin, Heidelberg, New York, 1969.

[FH] N. FUSCO AND J. HUTCHINSON, *$C^{1,\alpha}$ partial regularity of functions minimising quasiconvex integrals*, Manuscripta Math., 54 (1985), pp. 121–143.

[Ge] F. W. GEHRING, *The L^p–integrability of the partial derivatives of a quasiconformal map,* Acta Math., 130 (1973), pp. 265–277.

[G1] M. Giaquinta, *Multiple Integrals in the Calculus of Variations and Nonlinear Elliptic Systems*, Princeton University Press, Princeton, NJ, 1983.

[G2] M. Giaquinta, *Introduction to Regularity Theory for Nonlinear Elliptic Systems*, Birkhäuser, Basel, Boston, Berlin, 1993.

[GM] M. Giaquinta and G. Modica, *Partial regularity of minimizers of quasiconvex integrals*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 3 (1986), pp. 185–208.

[HW] P. Hartman and A. Wintner, *On uniform Dini conditions in the theory of linear partial differential equations of elliptic type,* Amer. J. Math., 77 (1955), pp. 329–354.

[Ko] J. Kovats, *Fully nonlinear elliptic equations and the Dini condition,* Comm. Partial Differtial Equations, 22 (1997), pp. 1911–1927.

[M1] C. B. Morrey, *Quasi-convexity and the lower semicontinuity of multiple integrals*, Pacific J. Math., 2 (1952), pp. 25–53.

[M2] C. B. Morrey, *Multiple Integrals in the Calculus of Variations*, Springer, Berlin, Heidelberg, New York, 1966.

[S1] L. Simon, *Lectures on Geometric Measure Theory*, Australian National University Press, Canberra, Australia, 1983.

[S2] L. Simon, *Theorems on Regularity and Singularity of Energy Minimizing Maps*, Birkhäuser, Basel, Boston, Berlin, 1996.

[Ta] I. Tamanini, *Regularity Results for Almost Minimal Oriented Hypersurfaces in* $\mathbb{R}^n$, Quad. Dipt. Mat. Uni. Lecce 1-1984, Università di Lecce, Italy, 1984.

SIAM J. MATH. ANAL.
Vol. 32, No. 3, pp. 688–706

PERIODIC STRUCTURE IN TWO-DIMENSIONAL RIEMANN PROBLEMS FOR HAMILTON–JACOBI EQUATIONS*

J. D. PINEZICH†

Abstract. This paper investigates the structure of two-dimensional Riemann problems for Hamilton–Jacobi equations. The solutions to such problems are fundamental building blocks for constructing solutions to more general problems, in particular, for numerical construction using methods such as front tracking. Here we prove the existence of a particular class of Riemann problem for which the viscosity solutions contain closed characteristic orbits, enclosing furthermore a periodic sonic structure, which in turn encloses a parabolic structure. The existence of such examples elucidates the difficulties encountered in designing construction methods for viscosity solutions to Riemann problems in dimension ≥ 2. This investigation was prompted by the discovery of numerical evidence of examples displaying an even richer internal structure.

Key words. Riemann problem, Hamilton–Jacobi equation, conservation law, sonic shock, delay differential equations, viscosity solutions

AMS subject classifications. 70H20, 35L67, 49L25, 35B05

PII. S0036141098342143

1. Introduction.

The central equation in this paper is the Hamilton–Jacobi equation

$$\mathcal{S}(v) - D\mathcal{S}(v) \cdot v + \mathcal{H}(D\mathcal{S}(v)) = 0, \tag{1}$$

$$\text{with } D\mathcal{S} \text{ specified at infinity.} \tag{2}$$

Here $D\mathcal{S}$ denotes the gradient of $\mathcal{S} : \mathbb{R}^2 \to \mathbb{R}$, and the Hamiltonian $\mathcal{H} : \mathbb{R}^2 \to \mathbb{R}$ is a supposed known function of $D\mathcal{S}$. Assuming only continuity of $\mathcal{H}$, (1), (2) has a unique viscosity solution $\mathcal{S}$ (see [9] based on [4, 5, 14]); nevertheless the construction of the solution is rarely straightforward. Exceptions occur when either $\mathcal{H}$ or data $D\mathcal{S}$ are convex, for which $\mathcal{S}$ is obtained from $\mathcal{H}$ and $D\mathcal{S}$ through the Legendre transform [3, 9]. Classification of solution types for problem (1), (2) is an open problem. We develop a geometric framework for understanding the structure of solutions and singularities arising therein. Within this framework we exhibit a class of problems for which the viscosity solution has a complex internal structure. The work here establishes tools for the construction of a variety of classes of solution types.

Solutions of (1), (2) are related to solutions of the Hamilton–Jacobi equation

$$D_t \phi + \mathcal{H}(D_z \phi) = 0 \tag{3}$$

for a function $\phi : \mathbb{R}^2 \times \mathbb{R}^+ \to \mathbb{R}$, with initial data at $t = 0$. Equation (3) occurs in many contexts as an evolution model, e.g., the propagation of wavefronts and the evolution of material interfaces as occurs under etching and deposition processes in chip manufacture [9, 10, 12, 24, 25]. Equation (3) is the integrated form of a

*Received by the editors July 21, 1998; accepted for publication (in revised form) July 28, 1999; published electronically October 31, 2000. This work was supported in part by the Army Research Office under grants DAAH0493G0334, FY9192, DAAH049510266, DAAH0049610123, DAAH049510414, and DAAH049510266 and the National Science Foundation under grant DMS9500568.

http://www.siam.org/journals/sima/32-3/34214.html

†Department of Applied Mathematics and Statistics, State University of New York, Stony Brook, NY 11794-3600 (pinezich@ams.sunysb.edu).

conservation law for $p = D_z\phi$: $D_t\, p + D_z\, \mathcal{H}(p) = 0$, and one expects discontinuities (e.g., shocks) to arise in the gradient of ϕ. Conservation laws generally suffer from a lack of uniqueness of solutions, and the notion of viscosity solutions guaranteeing uniqueness for Hamilton–Jacobi equations was introduced precisely for this reason.

A natural framework for understanding the evolution of singularities in solutions to (3) is to seek them in self-similar form: $\phi(z,t) = t\,\mathcal{S}(z/t)$. With $v = z/t$, (1) is the reduced equation for $\mathcal{S}$ [3, 9]. We refer to the boundary condition at infinity, induced by the directional derivatives of $\phi(z,0)$ at $z = 0$, as the *Riemann data*, the viscosity solution $\mathcal{S}$ as the *Riemann solution*, and (1), (2) as a *Riemann problem*.

The theory of Legendre transforms has its parallel in the construction of solutions to (3), i.e., closed forms for $\phi(z,t)$ in terms of $\mathcal{H}$ and the initial data. Such formulas were first obtained for one spatial dimension [20] with convex $\mathcal{H}$ or initial data and later under similar hypotheses extended in [2, 13, 18]. These ideas have been further extended to certain nonconvex cases [3]. Such results allow for the decomposition of an arbitrary function $\mathcal{H}$ as a sum of simpler ones and have led to Godunov-type algorithms for the numerical construction of viscosity solutions for (3) [1, 19, 21]. Such algorithms yield then, in principle, numerical algorithms for the construction of solutions to the Riemann problem (1), (2).

The front tracking [7, 8, 10, 12] approach to (3) is opposite in spirit, as it uses knowledge of Riemann solutions to propagate singularities in the solution and higher order methods for the propagation of the solution where it is smooth. Such algorithms are more efficient than regular solvers and there is considerable interest in constructing Riemann solutions and in understanding their structure. Although the local structure of Riemann solutions for Hamilton–Jacobi equations is well understood [9], a complete theory of their global structure and algorithms for their construction is lacking at the time of this writing. It is hoped that the example studied here, by providing insight to the structure of solutions, will further this goal, as in the case of Riemann problems for the related two-dimensional conservation laws [17, 26, 28].

Characteristics for (1) are straight lines and the method of characteristics constructs the proper solution to (1), (2) near infinity where characteristic velocities point radially inward. Within a compact set, piecewise characteristic closed paths can form, along which the Cauchy problem fails to be hyperbolic. The existence of such cases was conjectured earlier and in this paper we provide the first rigorous construction of this phenomenon.

A shock is called *sonic* if characteristics are tangent to it. Sonic shocks can be characteristic, and thus straight lines, or noncharacteristic, and thus curved.

DEFINITION 1. *A* sonic sequence *is a sequence V_i, $i = 0,\ldots,N-1$, of noncharacteristic sonic shocks, such that all characteristics incident upon shock V_i, $i = 1,\ldots,N-1$, leave shock V_{i-1} tangentially. If all characteristics incident upon V_0 leave shock V_{N-1} tangentially, the sonic sequence is said to be* periodic *with* period N.

The main result of this paper is the following theorem.

THEOREM 1. *There exist C^1 Hamiltonians and Riemann problems so that the associated Riemann solutions to the Hamilton–Jacobi equation* (1) *possess a period* 4 *sonic sequence.*

This theorem was motivated by the Riemann problem

$$\mathcal{H}(p) = \xi\eta + \frac{1}{2}(\eta^4 - \xi^4), \tag{4}$$

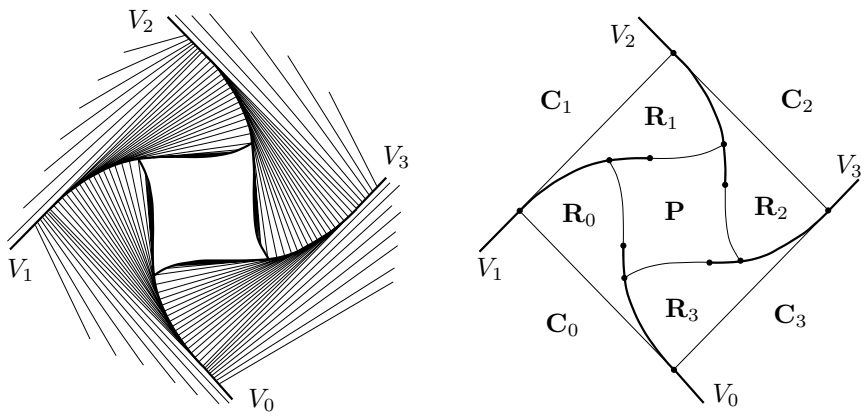


FIG. 1. *Conjectured structure of viscosity solution to* (4), (5). *Left*: *shocks* V_i *and characteristics. Right*: $\mathbf{C}_i$, $\mathbf{R}_i$, $\mathbf{P}$ *are domains of plane waves, sonic rarefactions, and a parabolic wave.*

where $p = (\xi, \eta)$, and Riemann data

$$D\mathcal{S}(\theta) = \begin{cases} \alpha_3 = (0,1) & \text{if} \quad -\pi/4 \le \theta < \pi/4, \\ \alpha_2 = (-1,0) & \text{if} \quad \pi/4 \le \theta < 3\pi/4, \\ \alpha_1 = (0,-1) & \text{if} \quad 3\pi/4 \le \theta < 5\pi/4, \\ \alpha_0 = (1,0) & \text{if} \quad -3\pi/4 \le \theta < -\pi/4 \end{cases} \tag{5}$$

posed by Tangerman and Kranzer, for which a numerical solution was found containing a period 4 sonic sequence [22, 23], shown in Figure 1. Regions $\mathbf{C}_i$ are the domains of plane waves (see below), on which the solution is linear; regions $\mathbf{R}_i$ are the domains of sonic rarefactions, on which the solution is a union of lines tangent to the curves V_i; and the region $\mathbf{P}$ is the domain of a parabolic wave, on which the solution is a saddle-type function.

To prove Theorem 1 we construct a somewhat simpler function $\mathcal{S}$ containing a period 4 sonic sequence, see Figure 2, and from it derive $\mathcal{H}$ and data $D\mathcal{S}$ such that $\mathcal{S}$ is a Riemann solution. Characteristics form two closed paths, one joining points B_i, the other joining points A_i, and the shock segments between these points form the sonic sequence. The main difficulty in the construction arises from a geometric constraint on continuity imposed by the periodicity: characteristic paths leaving a shock ultimately return to that shock. We remark that period 3 sonic sequences are trivial since a closed path of three characteristics lies on a straight line.

In section 2 we provide relevant background material. In section 3 we give an example for which the Riemann solution contains a closed characteristic path bounding a parabolic wave, later used in section 7. Section 4 describes the geometric structure of sonic rarefactions and in section 5 we derive properties of sonic sequences required for the construction of periodic sonic sequences in section 6. The proof of Theorem 1 is found in section 7, where we simultaneously construct $\mathcal{S}$, $\mathcal{H}$, and Riemann problems. The author would like to thank Tangerman for suggesting this problem and his insights toward its resolution.

2. Background. For Hamilton–Jacobi equations, Crandall, Evans, and Lions introduced the concept of a *viscosity solution* and demonstrated existence and unique-

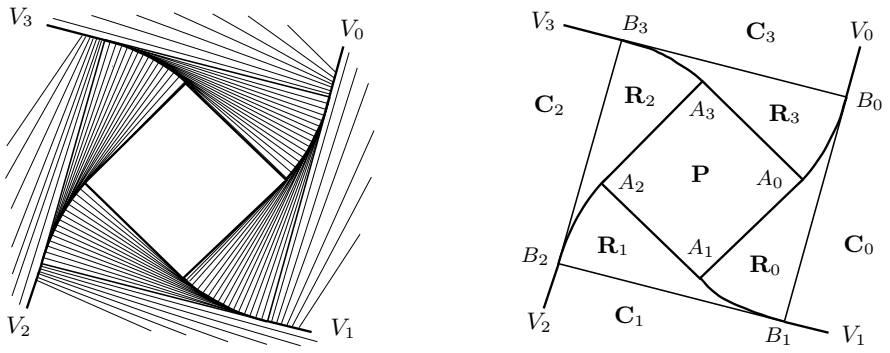


FIG. 2. *Solution with a period* 4 *sonic sequence, used in the proof of Theorem* 1. *The solution here is simpler than that in Figure* 1; *there, characteristics leave the parabolic wave.*

ness in this category [4]. $p \in \mathbb{R}^2$ is a *superderivative*, resp., *subderivative* [4, 9] of $\mathcal{S}$ at $v_0 \in \mathbb{R}^2$ if

$$\limsup_{v \to v_0} \frac{\mathcal{S}(v) - \mathcal{S}(v_0) - p \cdot (v - v_0)}{|v - v_0|} \le 0, \qquad \liminf_{v \to v_0} \frac{\mathcal{S}(v) - \mathcal{S}(v_0) - p \cdot (v - v_0)}{|v - v_0|} \ge 0.$$

Let $C^\wedge(v_0)$ be the set of superderivatives and $C^\vee(v_0)$ the set of subderivatives at v_0, and let $C^{\wedge\vee}(v_0) = C^\wedge(v_0) \cup C^\vee(v_0)$. $C^\wedge(v_0)$ and $C^\vee(v_0)$ are closed, convex, and possibly empty [9]. If $\mathcal{S}$ is differentiable at v_0, then the gradient $p = D\mathcal{S}(v_0)$ is the unique super- and subderivative. Conversely, if both $C^\wedge(v_0)$ and $C^\vee(v_0)$ are nonempty, then each consists of the same element, namely, the derivative of $\mathcal{S}$ at v_0. At a point v_0 on a shock, $C^{\wedge\vee}(v_0)$ is the convex hull of the gradients $p, q \in \mathbb{R}^2$ on each side of the shock. By continuity, $p - q \in \mathbb{R}^2$ is normal to the shock. If $p - q$ points to the side with $D\mathcal{S} = p$ (equivalently, $q - p$ points to side with $D\mathcal{S} = q$), the shock is a *convex edge* and $C^{\wedge\vee}(v_0) = C^\vee(v_0)$; otherwise it is a *concave edge* and $C^{\wedge\vee}(v_0) = C^\wedge(v_0)$.

$\mathcal{S}$ is a *viscosity solution* [4, 9] of (1) if for all $v \in \mathbb{R}^2$

$$\begin{aligned} &\mathcal{S}(v) - v \cdot p + \mathcal{H}(p) \le 0 \quad \text{for all } p \in C^\wedge(v) \\ \text{and} \quad &\mathcal{S}(v) - v \cdot p + \mathcal{H}(p) \ge 0 \quad \text{for all } p \in C^\vee(v). \end{aligned}$$

For any given Riemann data and continuous $\mathcal{H}$, (1), (2) has a unique Lipschitz continuous Riemann solution $\mathcal{S}$ [4, 9, 14].

With $p = D\mathcal{S}$ the characteristic equations for (1) are the system of ODEs [9, 15]

$$D_\tau p = 0, \qquad D_\tau v = -v + D_p \mathcal{H}(p), \qquad D_\tau \mathcal{S} = -v \cdot p + D_p \mathcal{H}(p) \cdot p.$$

Outside a compact set these equations are nonsingular and the Riemann data can be propagated inward by hyperbolic methods [9]. By the first of these equations, p, and thus $D_p\mathcal{H}(p)$, are constant along characteristics, hence characteristics are straight lines. They are directed toward points given by the *parabolic mapping* $L(p) := (v(p), \mathcal{S}(p)) : \mathbb{R}^2 \to \mathbb{R}^2 \times \mathbb{R}$ [9]:

$$v(p) = D_p \mathcal{H}(p), \qquad \mathcal{S}(p) = v(p) \cdot p - \mathcal{H}(p). \tag{6}$$

A characteristic path can terminate at a shock or, if the shock is sonic, continue along the outgoing tangential shock. By an *orbit* we mean a closed path of characteristics.

Define the graph of the mapping $p \to v_0 \cdot p - \mathcal{S}(v_0)$ over $C^{\wedge\vee}(v_0)$ as the *chord* of $C^{\wedge\vee}(v_0)$. The viscosity condition can be expressed as a generalization of the Oleĭnik condition [20].

Oleĭnik: *$\mathcal{S}$ is a Riemann solution if and only if the chord of $C^{\wedge}(v)$ lies on or above the graph of $\mathcal{H}$ and the chord of $C^{\vee}(v)$ lies on or below the graph of $\mathcal{H}$ for all $v \in \mathbb{R}^2$.*

As a consequence, characteristics satisfy the Lax condition [16] at a shock [9].

Lax: *Characteristic directions point toward the shock or are tangent to it.*

In this paper all Riemann solutions are piecewise smooth and consist of three wave-types classified by rank of the matrix $D^2\mathcal{S}(v)$ of second derivatives of $\mathcal{S}$ at a point v. Let E be an open connected set on which a Riemann solution is C^2. If $\operatorname{rank} D^2\mathcal{S}(v) = 2$ on E, the solution is called a *parabolic wave*. If $\operatorname{rank} D^2\mathcal{S}(v) = 1$ on E the solution is called a *rarefaction*. If $\operatorname{rank} D^2\mathcal{S}(v) = 0$ on E the solution is called a *plane wave*.

Throughout the paper we exploit the symmetry of (1). With $p = D\mathcal{S}$, (1) becomes

$$\mathcal{S}(v) - v \cdot p + \mathcal{H}(p) = 0,$$

and the change of notation $v \leftrightarrow p$, $\mathcal{S} \leftrightarrow \mathcal{H}$ results in the same equation. The plane on which $\mathcal{S}$ is defined is called the *v-plane*; the plane on which $\mathcal{H}$ is defined is called the *p-plane*.

3. Example with period 4 orbit. The solution to the Riemann problem with Hamiltonian $\mathcal{H}(p) = (\xi^2 - \eta^2)/2$, where $p = (\xi, \eta)$, and Riemann data $\{\zeta_i\}$ as in Figure 3 contains a period 4 orbit bounding a parabolic wave. Let $\zeta_i = (\xi_i, \eta_i)$. By (6), characteristics are directed toward points $A_i = (a_i, \mathcal{S}(a_i))$ given by $a_i = (\xi_i, -\eta_i)$, $\mathcal{S}(a_i) = (\xi_i^2 - \eta_i^2)/2$. $\mathbf{P}$ is the convex hull of the set $\{a_i\}$ and is the domain of the parabolic wave. For each i, the set $\mathbf{C}_i = a_i + \bigcup_{\kappa,\kappa' \geq 0} \kappa\,(a_{i-1} - a_i) + \kappa'\,(a_i - a_{i+1})$ is the domain of a plane wave with gradient ζ_i. These waves are joined by shocks V_i that propagate inward from infinity along straight lines.

This example furnishes us with a structure from which we can construct outward a solution containing a sonic sequence, as done in the following sections, and which is ultimately used to prove Theorem 1. The idea is to replace the plane waves with sonic rarefactions tangential to the plane waves at the parabolic boundary.

PROPOSITION 1. *The Riemann solution to the above Riemann problem is*

$$\mathcal{S}(v) = \begin{cases} (x^2 - y^2)/2, & v \in \mathbf{P}, \\ y + 1/2, & v \in \mathbf{C}_0, \\ -x - 1/2, & v \in \mathbf{C}_1, \\ -y + 1/2, & v \in \mathbf{C}_2, \\ x - 1/2, & v \in \mathbf{C}_3, \end{cases} \tag{7}$$

where $v = (x, y)$. The boundary of the parabolic wave is an orbit of period 4.

Proof. Where differentiable, $\mathcal{S}$ satisfies (1). On $\mathbf{P}$, $\mathcal{S}$ is a ruled surface with two distinct rulings that coincide with four characteristic shocks V_i from infinity which form an orbit bounding the parabolic wave. These shocks are also the boundaries of the plane waves and hence $\mathcal{S}$ is continuous. $\mathcal{H}$ is also a ruled surface and the chords of $C^{\wedge\vee}(v)$ along the shocks lie on the graph of $\mathcal{H}$ and hence the Oleĭnik condition is satisfied. □

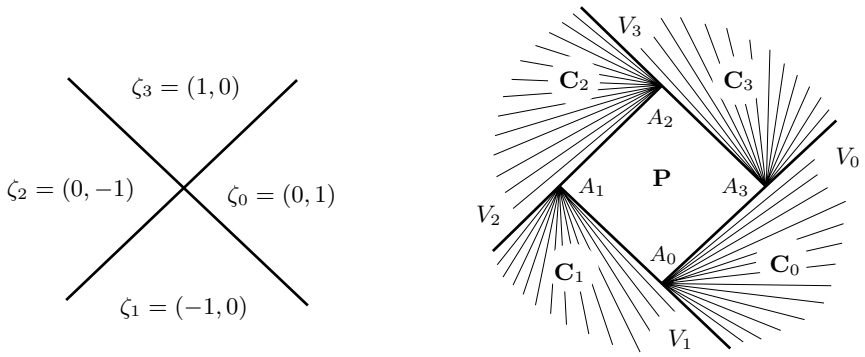


FIG. 3. *Riemann solution with orbit. Left: Riemann data* $\{\zeta_i\}$. *Right: shocks* V_i *and characteristics directed toward points* $\mathrm{A_i}$. $\mathbf{C}_i$, $\mathbf{P}$ *are domains of plane and parabolic waves, respectively.*

4. Structure of sonic rarefactions. Sonic rarefactions occur when characteristics leave a shock V tangentially. Since characteristics are straight lines, a sonic rarefaction is a type of ruled surface, called the *tangent surface* of the curve V. Given a curve V with tangent surface described by a function $\mathcal{S}$, we show that there exists a Hamiltonian $\mathcal{H}$ such that (1) is satisfied. The properties developed here, and in section 5 (sequences of sonic rarefactions), allow for the construction in section 6 of a period 4 sonic sequence, using an orbit as starting point (as from, e.g., the example in section 3). Before proceeding, we provide basic terminology.

A *parameterized curve* is a C^2 differentiable map $\Gamma : I \to \mathbb{R}^n$ $(n = 2, 3)$, where I is the unit interval $[0, 1]$. If $\dot{\Gamma}(t) \neq 0$ for all $t \in I$, the curve is *regular* (the notation $\dot{f}(t) = D_t f(t)$ indicates differentiation of f with respect to t). Two parameterized curves $\Gamma, \bar{\Gamma} : I \to \mathbb{R}^n$ are *equivalent* if there exists an orientation-preserving diffeomorphism $\phi : I \to I$ such that $\bar{\Gamma}(t) = \Gamma(\phi(t))$, $t \in I$. By the *oriented curve* Γ we mean the equivalence class of all parameterized curves equivalent to Γ. Let $v : I \to \mathbb{R}^2$ be an oriented curve which can be written in an orthogonal frame as $v(x) = (x, \varphi(x))$. If φ is strictly convex then v is called a *convex curve*.

For $z \in \mathbb{R}^2$ we denote as z^* the rotation of z about the origin by $+90°$. For vectors $w, z \in \mathbb{R}^2$ we can compute their determinant $\|w\ z\|$ as $w^* \cdot z$. For a convex curve $v : I \to \mathbb{R}^2$, the determinant $\|\dot{v}(t)\ \ddot{v}(t)\| \neq 0$ for all $t \in I$, and v lies entirely in a closed half plane which has boundary given by a tangent line of v. The interior of the intersection of all such half planes of v is called the *inside* of v. The *outside* of v is the closure of the complement of the inside of v. Note that the inside of a convex planar curve is convex.

We now describe the tangent surface to a curve; see Figure 4. Since our construction of the periodic sequence in section 6 is *outward* from an orbit, the choice of orientation in (8) is the natural one. Let $v : I \to \mathbb{R}^2$ be a regular curve and let $S : I \to \mathbb{R}$ be a C^2 function. Denote by $V : I \to \mathbb{R}^2 \times \mathbb{R}$ the curve defined by $V(t) = (v(t), S(t))$. Let $\mathbf{U} = \{(t, s) \mid 0 \leq t \leq 1, s > 0\}$ and consider the mapping $\Omega : \overline{\mathbf{U}} \to \mathbb{R}^2 \times \mathbb{R}$, where $\overline{\mathbf{U}}$ is the closure of $\mathbf{U}$, given by

$$\Omega(t, s) = V(t) - s\dot{V}(t). \tag{8}$$

DEFINITION 2. *The image* $\Omega(\overline{\mathbf{U}})$ *is called the* tangent surface *of* V. *The tangent half line given by* $\Omega(t, s)$, $s > 0$, *is called the* characteristic *at* $V(t)$.

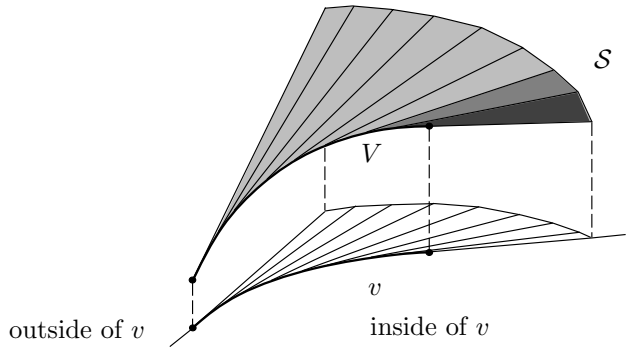


FIG. 4. *The structure of a sonic rarefaction is described by the tangent surface of a curve $V = (v, S)$. For convex v, $\mathcal{S}$ is a function representing the tangent surface. The inside of v is the intersection of half planes containing v.*

If v is convex, $\Omega(\overline{\mathbf{U}})$ is a graph of a C^1 function $\mathcal{S}$ with domain the tangent surface of v. We show in Proposition 2 that for some $\mathcal{H}$, $\mathcal{S}$ is a solution to (1). Of principal importance in the following lemma is the orthogonality of v and p, leading naturally to a notion of *duality*, and further exploited in developing properties of sonic sequences, and in the construction of our solution.

LEMMA 1. (a) *The gradient $D\mathcal{S}$ of $\mathcal{S}$ is constant along characteristics and equals*

$$D\mathcal{S}(\omega(t,s)) = \frac{-\ddot{S}(t)\dot{v}^*(t) + \dot{S}\ddot{v}^*(t)}{\|\ddot{v}(t)\ \dot{v}(t)\|} =: p(t). \tag{9}$$

(b) $\dot{p}(t) \cdot \dot{v}(t) = 0$, (c) $\dot{S}(t) = p(t) \cdot \dot{v}(t)$, $t \in I$.

Proof. (a) Write $\Omega(t,s) = (\omega(t,s), S(t,s))$, where $\omega(t,s) = v(t) - s\dot{v}(t)$ and $S(t,s) = S(t) - s\dot{S}(t)$. By the chain rule $D\mathcal{S}(\omega(t,s)) = DS(t,s)(D\omega(t,s))^{-1}$. Evaluating this we obtain (9), and since this is independent of s, the gradient is constant along characteristics. (b) Differentiate (9) and carry out the dot product. (c) Since $\omega(t,0) = v(t)$ we have $\dot{S}(t) = D_t\mathcal{S}(v(t)) = D\mathcal{S}(v(t)) \cdot \dot{v}(t) = p(t) \cdot \dot{v}(t)$. ∎

Note that (9) is well defined for equivalence classes of parameterized curves: suppose $\bar{V}(t) = V(\phi(t))$, then $\bar{p}(t) = p(\phi(t))$. If V is C^∞, then p is C^∞ and hence $\mathcal{S}$ is C^∞ for $s \neq 0$. For $s = 0$, i.e., along v, the (s, t)-coordinate system is singular and there $\mathcal{S}$ is usually only C^1.

For a given tangent surface $\mathcal{S}$ determined by a curve $V = (v, S)$, we have obtained an expression $p(t)$ for $D\mathcal{S}$ and shown that $\dot{p}(t) \perp \dot{v}(t)$. To an additive constant, the pair (v, p) is an equivalent descriptor of $\mathcal{S}$. Let $v, p : I \to \mathbb{R}^2$ satisfy Lemma 1(b), and $S : I \to \mathbb{R}$ satisfy Lemma 1(c). If $\mathcal{S}$ is the tangent surface of $V = (v, S)$, then $D\mathcal{S}(\omega(t,s)) = p(t)$. This result allows us to construct our solution in terms of pairs $(v, p) : I \to \mathbb{R}^2 \times \mathbb{R}^2$ in the v- and p-planes.

DEFINITION 3. *Let $v, p : I \to \mathbb{R}^2$ be curves such that $\dot{v}(t) \cdot \dot{p}(t) = 0$, $t \in I$. Then (v, p) is called an* orthogonal pair. *If v and p are also convex, then (v, p) is called a* dual pair. *Let $\mathcal{S}$ be a tangent surface with dual pair (v, p). Let H satisfy $\dot{H}(t) = v(t) \cdot \dot{p}(t)$ and $H(0) = v(0) \cdot p(0) - S(0)$, and let $\mathcal{H}$ be the tangent surface of $P(t) := (p(t), H(t))$. Then H, P, and $\mathcal{H}$ are said to be* dual *to S, V, and $\mathcal{S}$, respectively. Characteristics tangent to P (or p) are* dual *to those tangent to V (or v).*

PROPOSITION 2. *Let $\mathcal{H}$ be dual to $\mathcal{S}$.* (a) *The characteristics at $v(t)$ and $p(t)$ are orthogonal.* (b) *On the characteristic at $p(t)$, $D_p\mathcal{H} = v(t)$.* (c) *If $\mathcal{H}$ is dual to $\mathcal{S}$, then*

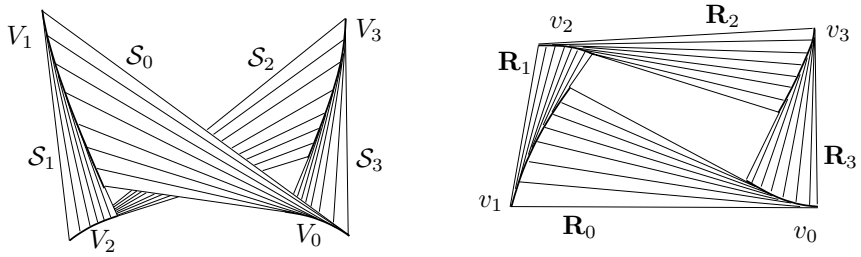


FIG. 5. *A periodic structure on the left, with annular domain on the right.*

$\mathcal{S}$ *satisfies* (1) *with Hamiltonian* $\mathcal{H}$.

Proof. (a) By Lemma 1(b).

(b) By Lemma 1(a) with $\mathcal{H}$, H, p in place of $\mathcal{S}$, S, v, respectively.

(c) Let z be on the characteristic at $v(t)$. Then $\mathcal{S}(z) = \mathcal{S}(v(t)) + p(t) \cdot (z - v(t))$. Writing $z = v(t) + (z - v(t))$, the left-hand side of (1) becomes $\mathcal{S}(z) - z \cdot D\mathcal{S}(z) + \mathcal{H}(D\mathcal{S}(z)) = \mathcal{S}(v(t)) - v(t) \cdot p(t) + \mathcal{H}(p(t))$. The derivative with respect to t of the right-hand side is zero since $D_p\mathcal{H}(p(t)) = v(t)$ by (b). By the duality of $\mathcal{S}$ and $\mathcal{H}$, $\mathcal{S}(v(0)) + \mathcal{H}(p(0)) = S(0) + H(0) = v(0) \cdot p(0)$, and thus $\mathcal{S}$ satisfies (1). ∎

5. Sequences of sonic rarefactions. If V_1 is a curve embedded in the tangent surface of a curve V_0, we can form a sequence of two tangent surfaces by truncating characteristics from V_0 along V_1. Generalizing this idea we can form sequences of length N, and if V_0 is embedded in the tangent surface of V_{N-1}, characteristic paths cycle periodically through the N curves. Here we derive properties of sequences of length $2, 3$, and 4, necessary for the construction carried out in section 6.

For $i = 0, 1, 2, 3$ let $\mathcal{S}_i$ be the tangent surfaces of curves $V_i = (v_i, S_i)$, such that characteristics tangent to V_i form an orientation-preserving diffeomorphism from V_i to V_{i+1} (here $V_{3+1} = V_0$); see Figure 5. Truncate characteristics from V_i at points on V_{i+1}; the domain of the resulting $\mathcal{S}_i$ has v_i and v_{i+1} in its boundary and is denoted by $\mathbf{R}_i$. If the interiors of the domains $\mathbf{R}_i$ and $\mathbf{R}_j$, $i \neq j$, are disjoint, the function $\mathcal{S}$ given by $\mathcal{S}(v) = \mathcal{S}_i(v)$, $v \in \mathbf{R}_i$, is called a *periodic structure*. We assume that the domain of the periodic structure is a topological annulus embedded in the plane and that the boundary curves of the annulus are orbits of the characteristic flow.

A periodic structure satisfying (1) for some $\mathcal{H}$ is a period 4 sonic sequence. However, a period 4 sonic sequence need not be a periodic structure: characteristics in a periodic sonic sequence can escape. The numerical solution [23, 22] motivating Theorem 1 contains such a sequence, joining a periodic structure to a parabolic wave.

DEFINITION 4. *Let* $\Gamma_i, \Gamma_j : I \to \mathbb{R}^n$, $i \neq j$ *be curves. By a* common parameterization *we mean that* $\Gamma_i(t)$ *and* $\Gamma_j(t)$ *are joined by a characteristic path for each* $t \in I$.

Note in the following lemma that (11) is independent of the curve v_{i+1}; this anticipates our method of construction in section 6. First construct V_0 and V_2 and from these derive V_1 and V_3 such that the sequence V_0, V_1, V_2, V_3 forms a periodic structure.

LEMMA 2. *Let* $\mathcal{S}_i, \mathcal{S}_{i+1}, \mathcal{S}_{i+2}$ *be tangent surfaces of curves* $V_i, V_{i+1}, V_{i+2} : I \to \mathbb{R}^2 \times \mathbb{R}$ *with common parameterization.* (a) $\mathcal{S}_i(v_{i+1}(t)) = \mathcal{S}_{i+1}(v_{i+1}(t))$ *if and only if*

$(p_i(t)-p_{i+1}(t))\cdot\dot v_{i+1}(t)=0$ *and* $\mathcal{S}_i(v_{i+1}(0))=\mathcal{S}_{i+1}(v_{i+1}(0))$. (b) *Suppose* $\mathcal{S}_i(v_{i+1}(t))=\mathcal{S}_{i+1}(v_{i+1}(t))$ *and* $\mathcal{S}_{i+1}(v_{i+2}(t))=\mathcal{S}_{i+2}(v_{i+2}(t))$, $t\in I$; *then*

$$p_{i+1}(t)\cdot(v_{i+2}(t)-v_{i+1}(t))=p_i(t)\cdot(v_{i+2}(t)-v_{i+1}(t)), \tag{10}$$

$$S_{i+2}(t)-S_i(t)=p_i(t)\cdot(v_{i+2}(t)-v_i(t)). \tag{11}$$

Proof. (a) This follows directly from continuity. (b) Since $\mathcal{S}_i$ and $\mathcal{S}_{i+1}$ join continuously along V_{i+1}, by (a) we have $(p_{i+1}(t)-p_i(t))\cdot\dot v_{i+1}(t)=0$, so that (10) holds. Since $\mathcal{S}_{i+1}$ and $\mathcal{S}_{i+2}$ join continuously along V_{i+2} we have $\mathcal{S}_{i+1}(v_{i+2}(t))=S_{i+2}(t)$ so that by (10) we obtain $S_{i+2}(t)-S_i(t)=p_{i+1}(t)\cdot(v_{i+2}(t)-v_{i+1}(t))+p_i(t)\cdot(v_{i+1}(t)-v_i(t))=p_i(t)\cdot(v_{i+2}(t)-v_i(t))$. □

Remark. In a period 3 orbit the three characteristics must lie in a plane, hence the gradient values on each characteristic lie along a straight line. By Lemma 2(a), the tangent vectors $\dot v_i(t)$ have direction orthogonal to this line; hence they, and therefore the characteristics, are all parallel. Thus period 3 sonic structures are trivial.

The periodic sonic sequence constructed in section 6 is bounded by orbital paths and in the following proposition we establish the geometry of such paths in the v- and p-planes.

Proposition 3. *Let periodic structure* $\mathcal{S}$ *have an orbit containing four points* $A_i=(a_i,\mathcal{S}_i(a_i))$ *with* $\{a_i\}$ *a strictly convex oriented quadrilateral. Then* $\{\alpha_i\}:=\{D\mathcal{S}_i(a_i)\}$ *is a strictly convex quadrilateral with opposite orientation.* (a) $(a_i-a_{i+1})\cdot(\alpha_i-\alpha_{i-1})=0$, (b) $(a_i-a_{i+2})\cdot(\alpha_i-\alpha_{i+2})=0$.

Proof. (a) follows from (10) of Lemma 2(b) and (b) from continuity along the orbit. An oriented quadrilateral with edges $\mathbf{e}_i$ is strictly convex if and only if $\|\mathbf{e}_i\ \mathbf{e}_{i+1}\|$ is of the same sign for all i, with sign given by orientation. Let $\mathbf{e}_i=a_i-a_{i+1}$ and $\mathbf{f}_i=\alpha_i-\alpha_{i-1}$. By (a), $\mathbf{f}_i=c_i\mathbf{e}_i^*$ for scalars c_i; thus $\alpha_{i+1}-\alpha_{i-1}=c_{i+1}\mathbf{e}_{i+1}^*+c_i\mathbf{e}_i^*$. Writing $a_{i+1}-a_{i-1}=\mathbf{e}_i+\mathbf{e}_{i-1}=-\mathbf{e}_{i+1}-\mathbf{e}_{i+2}$ and using $\mathbf{e}_{i+1}^*\cdot\mathbf{e}_i=\|\mathbf{e}_{i+1}\ \mathbf{e}_i\|$, we obtain from (b) $c_{i+1}/c_i=-\|\mathbf{e}_{i-1}\ \mathbf{e}_i\|/\|\mathbf{e}_{i+1}\ \mathbf{e}_{i+2}\|$. Since $\{a_i\}$ is strictly convex, the right-hand side is negative; hence c_{i+1} and c_i are of opposite sign. Thus $\|\mathbf{f}_i\ \mathbf{f}_{i+1}\|=c_ic_{i+1}\|\mathbf{e}_i\ \mathbf{e}_{i+1}\|$ is of sign opposite $\|\mathbf{e}_i\ \mathbf{e}_{i+1}\|$, demonstrating that $\{\alpha_i\}$ has orientation opposite that of $\{a_i\}$. □

Definition 5. *By* dual quadrilaterals *we mean strictly convex quadrilaterals* $\{a_i\}$ *and* $\{\alpha_i\}$ *of opposite orientation satisfying* (a) *and* (b) *in Proposition* 3.

Corollary 1. *The orbit of Proposition* 1 *has dual quadrilaterals* $\{a_i\}$ *and* $\{\zeta_{i+1}\}$.

Proof. The gradient of $\mathcal{S}$ on the characteristic joining A_i and A_{i+1} is ζ_{i+1}. □

Definition 6. *Let* $\mathcal{S}$ *be a periodic structure. The diffeomorphism* $\gamma: I\to I$ *defined as the first return from* v_0 *to* v_0 *by characteristic paths is called the* return map *of* $\mathcal{S}$.

We now derive a system of differential equations that the curves p_0 and p_2 in a periodic structure must satisfy. Note that these equations are independent of curves V_1 and V_3 (i.e., dual pairs (v_1,p_1) and (v_3,p_3)). Given appropriate curves v_0 and v_2 and return map γ, we can thus obtain curves p_0 and p_2, and we later show that the resulting (v_0,p_0) and (v_2,p_2) are dual pairs. These dual pairs allow for the construction of curves V_0 and V_2, and associated tangent surfaces, unique up to additive constants. From these, the curves V_1 and V_3 (hence the periodic structure) are derived.

Theorem 2. *Let characteristic paths in periodic structure* $\mathcal{S}$ *be given by the diagram*

$$\begin{pmatrix} v_0(t)\\ p_0(t)\end{pmatrix}\to\begin{pmatrix} v_1(t)\\ p_1(t)\end{pmatrix}\to\begin{pmatrix} v_2(t)\\ p_2(t)\end{pmatrix}\to\begin{pmatrix} v_3(t)\\ p_3(t)\end{pmatrix}\to\begin{pmatrix} v_0(\gamma(t))\\ p_0(\gamma(t))\end{pmatrix}. \tag{12}$$

Then p_0 and p_2 satisfy

$$\dot{p}_0(t) = \frac{(p_2(t) - p_0(t)) \cdot \dot{v}_2(t)}{\|\dot{v}_0(t)\ \ v_2(t) - v_0(t)\|} \dot{v}_0^*(t), \tag{13}$$

$$\dot{p}_2(t) = \frac{(p_2(t) - p_0(\gamma(t))) \cdot \dot{v}_0(\gamma(t))}{\|\dot{v}_2(t)\ \ v_2(t) - v_0(\gamma(t))\|} \dot{v}_2^*(t). \tag{14}$$

Proof. Applying (11) of Lemma 2(b) to $V_0(t), V_1(t), V_2(t)$ and $V_2(t), V_3(t), V_0(\gamma(t))$ yields $S_2(t) - S_0(t) = p_0(t) \cdot (v_2(t) - v_0(t))$ and $S_0(\gamma(t)) - S_2(t) = p_2(t) \cdot (v_0(\gamma(t)) - v_2(t))$. Taking derivatives and using $\dot{S}_i(t) = p_i(t) \cdot \dot{v}_i(t)$ from Lemma 1(c) gives

$$(p_2(t) - p_0(t)) \cdot \dot{v}_2(t) = (v_2(t) - v_0(t)) \cdot \dot{p}_0(t), \tag{15}$$

$$(p_2(t) - p_0(\gamma(t)) \cdot \dot{v}_0(\gamma(t)) = (v_2(t) - v_0(\gamma(t))) \cdot \dot{p}_2(t). \tag{16}$$

Since v_i is convex by assumption, $\dot{v}_i(t) \neq 0$, and Lemma 1(b) implies that $\dot{p}_i(t)$ has the form $N_i(t)\dot{v}_i^*(t)$. The coefficients $N_0(t)$ and $N_2(t)$ are found from (15) and (16) resp., and yield (13), (14). ∎

If $\gamma(t) \leq t$, (13) and (14) are known as a *delay* system with *bounded delay* and have properties similar to ordinary differential equations [6, 11]. If the denominators are nonzero, the system (13), (14) is globally Lipschitz in $p_i(t)$ and by the methods of Chapters 25 and 26 of [6] and Chapter 2 of [11] has a unique solution on I which is a C^2 function of initial conditions, parameters, and t. Suppose $\bar{v}_0(t) = v_0(\phi(t))$ (inducing $\bar{v}_2(t) = v_2(\phi(t))$), then $\bar{p}_0$, $\bar{p}_2$ satisfy $\bar{p}_i(t) = p_i(\phi(t))$, $i = 0, 2$, and therefore (13), (14) is well defined for equivalence classes of parameterized curves.

6. The periodic structure. We first construct dual pairs for $i = 0, 2$ and from these derive dual pairs for $i = 1, 3$ completing the periodic structure $\mathcal{S}$. The dual pairs generate a dual periodic structure $\mathcal{O}$ representing chords of super- and subderivatives of $\mathcal{S}$.

6.1. Construction of dual pairs. Let $\{a_i\}$ and $\{\alpha_i\}$ be square dual quadrilaterals centered at the origin and define vectors $\mathbf{e}_i = a_i - a_{i+1}$ and sectors $U_i = \bigcup_{\kappa,\kappa'>0}(\kappa\ \mathbf{e}_i + \kappa'\ \mathbf{e}_{i-1})$; see Figure 6. Each sector is of angle 90°, U_i is adjacent to U_{i+1}, and $a_i \in U_{i+1}$. Let $v_0, v_2 : I \to \mathbb{R}^2$ be C^∞ convex curves such that

(A1) $v_i(0) = a_i$ and v_i is in U_{i+1};

(A2) $\dot{v}_i(0)$ is in the direction $\mathbf{e}_i$ and $\dot{v}_i(t) \in U_i$ for $t > 0$.

Since $v_i(0) = a_i$ is in the open set U_{i+1} such curves exist. Let the correspondence between points on v_0 and v_2 through characteristic paths through v_1 be given by the orientation-preserving diffeomorphism $\phi_{012} : v_0 \to v_2$ and the correspondence between points on v_0 and v_2 through characteristic paths through v_3 be given by the orientation-preserving diffeomorphism $\phi_{230} : v_2 \to v_0$. Let v_0 and v_2 have a common parameterization given by paths of ϕ_{012}, i.e., $v_2(t) = \phi_{012}(v_0(t))$. Assume that $\gamma : I \to I$, determined by the composition $\phi_{230} \circ \phi_{012}$, satisfies $\gamma(t) \leq t$, $t \in I$.

The following lemma is implied by assumptions (A1) and (A2).

LEMMA 3. *v_0 is in the inside of v_2, and vice versa.*

Analogous to $\mathbf{e}_i$ and U_i we define vectors $\mathbf{f}_i = \alpha_i - \alpha_{i-1}$ and sectors $\Upsilon_i = \bigcup_{\kappa,\kappa'>0}(\kappa\ \mathbf{f}_{i+1} + \kappa'\ \mathbf{f}_i)$. Since $\{\alpha_i\}$ is dual to $\{a_i\}$, $\mathbf{f}_i = c_i\mathbf{e}_i^*$ with the scalars c_i alternating in sign (as in Proposition 3). Hence Υ_i is equal to either U_{i-1} or U_{i+1}.

PROPOSITION 4. *Let p_0, p_2 be the solution to the system* (13), (14) *of Theorem* 2 *with $p_0(0) = \alpha_0$ and $p_2(0) = \alpha_2$. Then $\dot{p}_0(0)$ has direction $\mathbf{f}_0$ and $\dot{p}_2(0)$ has direction*

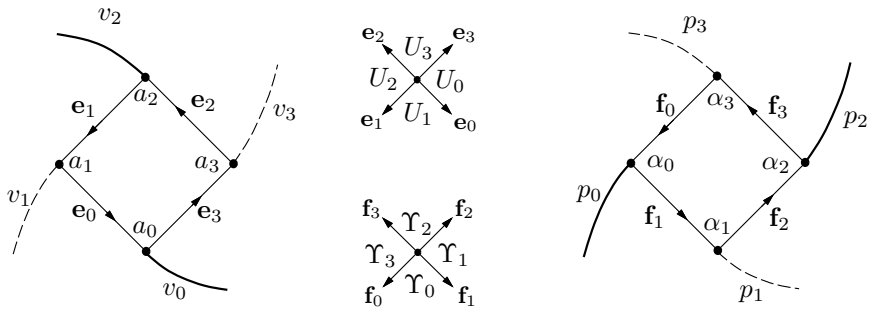


FIG. 6. *Specification of data*: *dual quadrilaterals* $\{a_i\}$, $\{\alpha_i\}$, *curves* v_0, v_2, *and diffeomorphisms* $\phi_{012} : v_0 \to v_2$, *and* $\phi_{230} : v_2 \to v_0$ *such that* $\gamma(t) \le t$. *Curve initial directions* $\mathbf{e}_i$, $\mathbf{f}_i$ *define sectors* U_i, Υ_i.

$\mathbf{f}_2$. *The curves* $V_0 = (v_0, S_0)$ *and* $V_2 = (v_2, S_2)$, *where* $\dot{S}_i(t) = p_i(t) \cdot \dot{v}_i(t)$, *satisfy*

$$S_2(t) - S_0(t) = p_0(t) \cdot (v_2(t)) - v_0(t)), \tag{17}$$

$$S_0(\gamma(t)) - S_2(t) = p_2(t) \cdot (v_0(\gamma(t)) - v_2(t)) \tag{18}$$

provided $S_2(0) - S_0(0) = \alpha_0 \cdot (a_2 - a_0) = \alpha_2 \cdot (a_2 - a_0)$.

Proof. By assumption (A2), $\dot{v}_i(0) = \epsilon_i \mathbf{e}_i$, $\epsilon_i > 0$, and hence $\dot{p}_i(0) = -\epsilon_{i+2} c_i c_{i+1} \|\mathbf{e}_{i+1}\ \mathbf{e}_{i+2}\| / \|\mathbf{e}_i\ \mathbf{e}_{i+1}\| \mathbf{f}_i$, and $\dot{p}_i(0)$ has direction $\mathbf{f}_i$. From (13) we obtain $(v_2(t) - v_0(t)) \cdot \dot{p}_0(t) = (p_2(t) - p_0(t)) \cdot \dot{v}_2(t)$. Subtracting $p_0(t) \cdot \dot{v}_0(t) = \dot{S}_0(t)$ from both sides yields $D_t[(v_2(t) - v_0(t)) \cdot p_0(t)] = D_t[S_2(t) - S_0(t)]$. Integrating from 0 to t gives (17). A similar argument establishes (18). □

Recall that if (v, p) is a dual pair, then the tangent surface of curve $V = (v, S)$, with $\dot{S}(t) = p(t) \cdot \dot{v}(t)$, has gradient $p(t)$ along the characteristic from $V(t)$ (Lemma 1). If we are given plane curves v_0, v_2, and a return map γ, by Proposition 4 we can construct curves V_0, V_2 such that the point $V_2(t)$ lies in the plane tangent to the tangent surface of V_0 at $V_0(t)$, and the point $V_0(\gamma(t))$ lies in the plane tangent to the tangent surface of V_2 at $V_2(t)$. This is precisely the geometry needed to construct curves V_1 and V_3.

Before proceeding with the construction we need to show that the curves p_0 and p_2 are regular. It is sufficient to show that the numerators in the coefficients N_0 and N_2 of (13) and (14) are nonzero. We first prove this for the special case where γ is the identity. Let $q(t) = p_2(t) - p_0(t)$. Subtracting (13) from (14) shows that q satisfies a linear equation $\dot{q}(t) = A(t)q(t)$ for some matrix $A(t)$. Since $q(0) = \alpha_2 - \alpha_0 \neq 0$, the solution q is never zero. Adding (17) to (18) in Proposition 4 shows that

$$(p_2(t) - p_0(t)) \cdot (v_2(t) - v_0(t)) = 0 \tag{19}$$

at all fixed points of γ. Thus $q(t) \perp (v_2(t) - v_0(t))$ for all $t \in I$. By assumption (A1) $v_2(t) - v_0(t) \in U_3$, and since $q(0) = \alpha_2 - \alpha_0 \in \Upsilon_1$ and $q(t) \neq 0$, the trace of q is in Υ_1. This proves that the numerators in N_0 and N_2 are never zero since by assumption (A2), $\dot{v}_i(t) \in \overline{U}_i$, and $\overline{U}_i$ is equal to either $\overline{\Upsilon}_1$ or $\overline{\Upsilon}_3$.

To prove regularity for general γ, we introduce the return map $\gamma_\mu : I \to I$, $\mu \in I$, defined by $\gamma_\mu(t) = (1 - \mu)t + \mu\gamma(t)$ which interpolates between γ and the identity γ_0. Since $\gamma_\mu(t) \le t$, Proposition 4 holds with γ_μ in place of γ; we denote the solutions as $p_i(t, \mu)$, noting that they are continuous functions of μ, and prove that they are regular for $\mu \in I$.

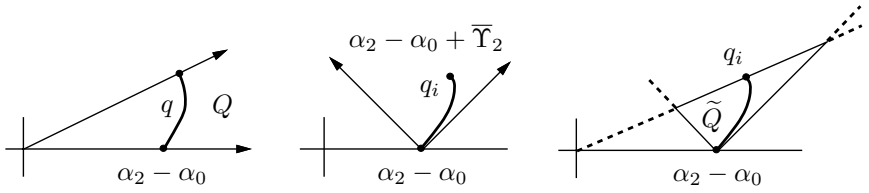


FIG. 7. *This figure pertains to the proof of Lemma* 4. *Q is the sector generated by q. The curves q_i are in Q and in $\alpha_2 - \alpha_0 + \overline{\Upsilon}_2$ and hence in their intersection, denoted $\widetilde{Q}$.*

LEMMA 4. *The curves p_0 and p_2 resulting from Proposition* 4 *are regular. p_2 is in the inside of p_0 and vice versa.*

Proof. Denote the now μ-dependent coefficients of (13) and (14) by $N_0(t,\mu)$ and $N_2(t,\mu)$, resp., and define the curves

$$q_0(t,\mu) = p_2(t,\mu) - p_0(t,\mu), \qquad q_2(t,\mu) = p_2(t,\mu) - p_0(\gamma_\mu(t),\mu).$$

It is enough to show that the numerators $q_0(t,\mu)\cdot \dot{v}_2(t)$ in N_0 and $q_2(t,\mu)\cdot \dot{v}_0(t)$ in N_2 are nonzero and we do this by proving that $q_i(t,\mu) \in \Upsilon_1$, $i = 0, 2$, for $(t,\mu) \in I \times I$.

Let $J = \{\mu \in I : q_i(t,\mu) \in \Upsilon_1; t \in I, i = 0, 2\}$. Since γ_μ is the identity for $\mu = 0$, and $q_i(t,0) = q(t) \in \Upsilon_1$ for $t \in I$, J is not empty. Since Υ_1 is open, J is open. Define

$$Q = \bigcup_{\kappa \geq 0, t \in I} \kappa\, q(t) \qquad \widetilde{Q} = Q \bigcap (\alpha_2 - \alpha_0 + \overline{\Upsilon}_2),$$

see Figure 7, and note that $\widetilde{Q}$ is a closed subset of Υ_1. We show that $\mu \in J$ implies $q_i(t,\mu) \in \widetilde{Q}$ for $t \in I$, from which it follows that J is closed; hence $J = I$.

Let $\mu \in J$. By assumption (A2), $\dot{v}_0(t) \in \overline{U}_0$ and $\dot{v}_2(t) \in \overline{U}_2$ for $t \in I$. Since $q_i(t,\mu) \in \Upsilon_1$, and Υ_1 is equal to either U_0 or U_2, the numerators in N_0 and N_2 are nonzero and of constant but opposite sign for $t \in I$. By Lemma 3, v_0 is in the inside of v_2 and vice versa; hence the denominators are also of constant but opposite sign, and therefore the coefficients $N_i(t,\mu)$ are nonzero and agree in sign on I. By Proposition 4, $\dot{p}_i(0,\mu)$ is in the direction $\alpha_i - \alpha_{i-1} = \mathbf{f}_i \in \overline{\Upsilon}_i$ and hence the signs are such that

$$\dot{p}_i(t,\mu) = N_i(t,\mu)\dot{v}_i(t) \in \overline{\Upsilon}_i. \tag{20}$$

Derivatives of q_i with respect to t are $\dot{q}_0(t,\mu) = \dot{p}_2(t,\mu) - \dot{p}_0(t,\mu)$ and $\dot{q}_2(t,\mu) = \dot{p}_2(t,\mu) - \dot{p}_0(\gamma_\mu(t),\mu)$, and therefore $\dot{q}_i(t,\mu) \in \overline{\Upsilon}_2$. From this we infer

$$q_i(t,\mu) \in q_i(\tau,\mu) + \overline{\Upsilon}_2 \tag{21}$$

for $t, \tau \in I$, $t \geq \tau$. In particular, $q_i(t,\mu) \in \alpha_2 - \alpha_0 + \overline{\Upsilon}_2$ for $t \in I$.

Now suppose there exists a point $q_i(\tau,\mu)$ not in $\widetilde{Q}$. Then by (21), $q_i(1,\mu)$ is not in $\widetilde{Q}$. But this is a contradiction: $t = 1$ is a fixed point of γ_μ, and by (19) must lie on a boundary line of Q, hence on a boundary line of $\widetilde{Q}$. Therefore $q_i(t,\mu) \in \widetilde{Q}$, $t \in I, i = 0, 2$. Thus J is closed and $J = I$. In particular, for $\mu = 1$, the numerators in N_0 and N_2 are nonzero. This proves that the curves p_0 and p_2 are regular.

By (20) we have $p_i(t) \in \alpha_i + \Upsilon_i, t \in I$. Since q_i is in Υ_1, we also have $p_i(t) \in \alpha_{i+1} + \Upsilon_{i-1}, t \in I$. Therefore p_i is in $(\alpha_i + \overline{\Upsilon}_i) \bigcap (\alpha_{i+1} + \Upsilon_{i-1})$, and it follows that p_2 is in the inside of p_0 and vice versa. ∎

THEOREM 3. *Let v_0, v_2 satisfy assumptions* (A1), (A2)*, and let p_0, p_2 be as in Proposition* 4*. Then (v_0, p_0) and (v_2, p_2) are dual pairs.*

Proof. For $i = 0, 2$, $\dot{p}_i(t) \cdot \dot{v}_i(t) = 0$, $t \in I$; thus (v_i, p_i) are orthogonal pairs. $\ddot{p}_i(t) = \dot{N}_i(t)\dot{v}_i^*(t) + N_i(t)\ddot{v}_i^*(t)$ so that $\|\dot{p}_i(t)\ \ddot{p}_i(t)\| = N_i^2(t)\|\dot{v}_i(t)\ \ddot{v}_i(t)\|$. Since p_i is regular by Lemma 4, $N_i(t) \neq 0$, and since v_i is convex the right-hand side is nonzero. Therefore p_i is convex and (v_i, p_i) are dual pairs. ∎

Remark. The assumption $\gamma(t) \leq t$ corresponds to characteristic paths that "spiral inward." The construction above has a natural dual for the case $\gamma(t) \geq t$, i.e., characteristic paths "spiral outward." Namely, we specify curves p_0, p_2 in place of v_0, v_2, and derive a system dual to (13), (14) for which v_0, v_2 are unknowns, and for which $\dot{v}_2(t)$ depends upon $v_0(\gamma^{-1}(t))$. Since $\gamma^{-1}(t) \leq t$, the dual system is a delay system and the method employed above is valid (interchanging v and p).

6.2. Completion of periodic structure. We now construct curves V_1 and V_3 which complete the periodic structure. We first obtain v_1 and v_3 as the solutions to differential equations, and then curves $V_1 = (v_1, S_1)$ and $V_3 = (v_3, S_3)$ by setting $S_1(t) = \mathcal{S}_0(v_1(t))$ and $S_3(t) = \mathcal{S}_2(v_3(t))$. The geometric relationship established in Proposition 4 provides that V_2 will lie in the tangent surface of V_1, and V_0 in the tangent surface of V_3, as shown below in Proposition 6.

In the v-plane, the periodic structure is captured by the diagram $v_0(t) \to v_1(t) \to v_2(t) \to v_3(t) \to v_0(\gamma(t))$. Since we are looking for sonic curves,

$$v_1(t) - v_0(t) = -l_0(t)\dot{v}_0(t), \tag{22}$$

$$v_2(t) - v_1(t) = -l_1(t)\dot{v}_1(t), \tag{23}$$

$$v_3(t) - v_2(t) = -l_2(t)\dot{v}_2(t), \tag{24}$$

$$v_0(\gamma(t)) - v_3(t) = -l_3(t)\dot{v}_3(t), \tag{25}$$

where l_i are positive scalar functions. From (22) we find two expressions for $l_0(t)$, one directly, the other upon differentiation:

$$\frac{\|\ddot{v}_0(t)\ \ v_1(t) - v_0(t)\|}{\|\dot{v}_0(t)\ \ddot{v}_0(t)\|} = l_0(t) = -\frac{\|\dot{v}_0(t)\ \dot{v}_1(t)\|}{\|\dot{v}_0(t)\ \ddot{v}_0(t)\|}. \tag{26}$$

From (23) we have $\dot{v}_1(t) = -(v_2(t) - v_1(t))/l_1(t)$, and inserting this to the right-hand expression for $l_0(t)$ in (26) and then using the left-hand expression yield

$$l_1(t) = -\frac{\|\dot{v}_0(t)\ \ v_2(t) - v_1(t)\|}{\|\ddot{v}_0(t)\ \ v_1(t) - v_0(t))\|}.$$

From (22), $\|\dot{v}_0(t)\ \ v_2(t) - v_1(t)\| = \|\dot{v}_0(t)\ \ v_2(t) - v_0(t)\|$; thus v_1 satisfies the differential equation

$$\dot{v}_1(t) = \frac{\|v_1(t) - v_0(t)\ \ \ddot{v}_0(t)\|}{\|\dot{v}_0(t)\ \ v_0(t) - v_2(t)\|}\ (v_1(t) - v_2(t)). \tag{27}$$

A similar equation holds for v_3:

$$\dot{v}_3(t) = \frac{\|v_3(t) - v_2(t)\ \ \ddot{v}_2(t)\|}{\|\dot{v}_2(t)\ \ v_2(t) - v_0(\gamma(t))\|}\ (v_3(t) - v_0(\gamma(t))). \tag{28}$$

PROPOSITION 5. *Let $i = 1$ (or 3) and let v_i be the unique solution of* (27) (*or* (28) *for $i = 3$) with initial condition $v_i(0) = a_i$. Then v_i is a C^∞ convex curve in $a_i + \overline{U}_i$. For $i = 0, 1, 2, 3$ the functions l_i are positive and v_i is inside v_{i+1}.*

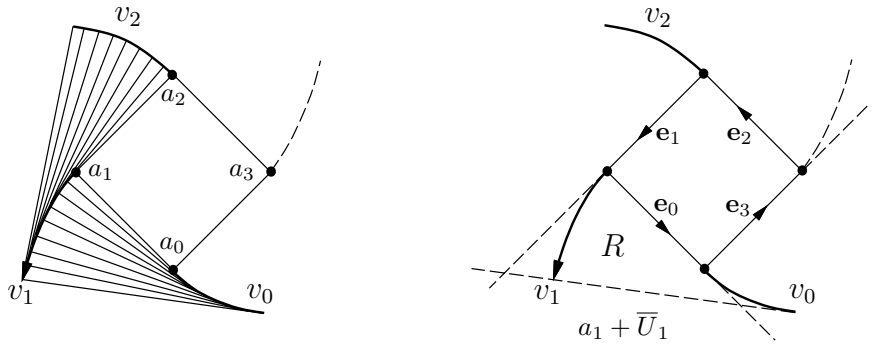


FIG. 8. *The geometrical relationship between v_0, v_1, and v_2. Left: characteristics are tangent to v_0 and v_1. Right: R is the intersection of the tangent surface of v_0 and $a_1 + \overline{U}_1$. v_1 is in R.*

Proof. Let $i = 1$ and consider the intersection R of the sector $a_1 + \overline{U}_1$ with the tangent surface of v_0; see Figure 8. R is a bounded set and since v_2 is inside v_0 (Lemma 3), (27) satisfies a Lipschitz condition on $I \times R$. The relations (22), (23) hold for the initial condition $v_1(0) = a_1 \in R$, and it follows that they hold for the unique C^∞ solution v_1 of (27); thus v_1 is in the tangent surface of v_0, and $l_0(t) > 0$. By assumptions (A1), (A2) the coefficient of $v_1(t) - v_2(t)$ in (27) is positive and thus v_1 is regular and $\dot{v}_1(t)$ has direction $v_1(t) - v_2(t) \in \overline{U}_1$. This implies v_1 is in $a_1 + \overline{U}_1$ and hence in R, and $l_1(t) > 0$. From (23) we compute $l_1(t)\|\dot{v}_1(t)\, \ddot{v}_1(t)\| = -\|\dot{v}_1(t)\, \dot{v}_2(t)\| \neq 0$ by assumptions (A1), (A2), and thus v_1 is convex. Similar results hold for $i = 3$. Finally, since $v_i \in a_i + \overline{U}_i$, v_i is inside v_{i+1} for all i. ∎

We remark that (27), (28) are well defined on equivalence classes of parameterized curves.

Since $l_i(t) > 0$, $v_{i+1}(t)$ is on the characteristic from $v_i(t)$ and we truncate characteristics from v_i along the curve v_{i+1}. Thus the closed set $\mathbf{R}_i$, defined by $\mathbf{R}_i = \bigcup_{t\in I} \bigcup_{0\le s\le l_i(t)} (v_i(t) - s\dot{v}_i(t))$, is the union of (truncated) characteristics tangent to v_i (see Figure 9). We show that the interiors of $\mathbf{R}_i$ and $\mathbf{R}_j$, $i \neq j$, are disjoint. For $j = i + 1$, $\mathbf{R}_i$ is in the inside of v_{i+1} and $\mathbf{R}_{i+1}$ is in the outside of v_{i+1}, hence their intersection is v_{i+1} and this is in the boundary of both. For $j = i + 2$, note that v_i and v_{i+1} are both in the inside of v_{i+2}. Since the inside of a curve is convex, $\mathbf{R}_i$ is inside v_{i+2} and disjoint from $\mathbf{R}_{i+2}$. Thus the union of the sets $\mathbf{R}_i$, $i = 0, 1, 2, 3$, forms a topological annulus, on which we define the periodic structure as follows.

Dual pair (v_0, p_0) generates a curve $V_0 = (v_0, S_0)$ which in turn generates a tangent surface described by a function $\mathcal{S}_0$ defined on $\mathbf{R}_0$. Similarly, dual pair (v_2, p_2) generates a curve V_2 and a tangent surface described by a function $\mathcal{S}_2$ defined on $\mathbf{R}_2$. The curves V_0, V_2, and hence the tangent surfaces, are unique up to additive constants and we fix these such that $S_2(0) - S_0(0) = \alpha_0 \cdot (a_2 - a_0)$. Curve v_1 is in the domain of $\mathcal{S}_0$ and v_3 is in the domain of $\mathcal{S}_2$, and we define $S_1(t) = \mathcal{S}_0(v_1(t))$ and $S_3(t) = \mathcal{S}_2(v_3(t))$. The curves $V_1(t) := (v_1(t), S_1(t))$ and $V_3(t) := (v_3(t), S_3(t))$ generate unique tangent surfaces given by functions $\mathcal{S}_1$ on $\mathbf{R}_1$ and $\mathcal{S}_3$ on $\mathbf{R}_3$. The four tangent surfaces define a periodic structure $\mathcal{S}$ on the topological annulus. Note that by Lemma 1(a) we obtain curves p_1, p_3 for the gradients $D\mathcal{S}_1$, $D\mathcal{S}_3$.

PROPOSITION 6. $\mathcal{S}_{i-1}(v_i(t)) = \mathcal{S}_i(v_i(t))$.

Proof. By construction the claim is true for $i = 1, 3$. Let $i = 0$, the case $i = 2$ is proved analogously. By the choice of additive constants, (17), (18) of Proposi-

tion 4 hold. From Proposition 5, $V_3(t)$ lies on the characteristic from $V_2(t)$ on which $D\mathcal{S}_2(t) = p_2(t)$. Thus $S_3(t) - S_2(t) = p_2(t) \cdot (v_3(t) - v_2(t))$. Subtracting this from (18) yields $S_0(\gamma(t)) - S_3(t) = p_2(t) \cdot (v_0(\gamma(t)) - v_3(t)) = p_3(t) \cdot (v_0(\gamma(t)) - v_3(t))$, where the second equality follows from $(p_2(t) - p_3(t)) \cdot \dot{v}_3(t) = 0$ of Lemma 2(a) and $\|\dot{v}_3(t)\ v_0(\gamma(t)) - v_3(t)\| = 0$. But then $\mathcal{S}_0(v_0(\gamma(t))) = S_0(\gamma(t)) = S_3(t) + p_3(t) \cdot (v_0(\gamma(t)) - v_3(t)) = \mathcal{S}_3(v_0(\gamma(t)))$. □

THEOREM 4. *Let $\{a_i\}$, $\{\alpha_i\}$ be square dual quadrilaterals centered at the origin, and let curves v_0, v_2 satisfy assumptions* (A1), (A2)*, and diffeomorphisms $\phi_{012} : v_0 \to v_2$, $\phi_{230} : v_2 \to v_0$ have return map $\gamma(t) \le t$. Then $\mathcal{S}$ is a periodic structure, unique up to an additive constant.*

Proof. By Proposition 6, $\mathcal{S}$ is continuous and so characteristics tangent to V_i form a map from V_i to V_{i+1}. Since the maps ϕ_{012} and ϕ_{230} are orientation-preserving diffeomorphisms and the curves V_i are regular, the characteristic maps from V_i to V_{i+1} are orientation-preserving diffeomorphisms. Hence $\mathcal{S}$ is a periodic structure. S_0 and S_2 are unique up to an additive constant and $\mathcal{S}$ is unique up to this constant also. □

Since $V_{i+1}(0)$ is on the characteristic from $V_i(0)$ for each i, these points are joined by an orbit. Algebraically we find $p_1(0) = \alpha_1$ and $p_3(0) = \alpha_3$. Characteristics joining points $V_{i+1}(1)$ also form an orbit with dual quadrilaterals $\{b_i\} := \{v_i(1)\}$ and $\{\beta_i\} := \{p_i(1)\}$.

6.3. Dual periodic structure completed. We continue our development by showing that (v_1, p_1) and (v_3, p_3) are dual pairs. This allows us to define a *dual* periodic structure $\mathcal{O}$, composed of tangent surfaces $\mathcal{H}_i$ dual to $\mathcal{S}_i$. The dual characteristics comprising the dual periodic structure correspond to the chords of super- and subderivatives $C^{\wedge\vee}(v_i)$ of the periodic structure $\mathcal{S}$. We use the dual periodic structure to show that $\mathcal{S}$ satisfies the Oleĭnik condition for the Hamiltonian $\mathcal{H}$ constructed in section 7.

Since adjacent surfaces $\mathcal{S}_i$ join continuously, from Lemma 2(a), the jump in $D\mathcal{S}$ across V_i is orthogonal to $\dot{v}_i(t)$, hence parallel to $\dot{p}_i(t)$. If $\dot{p}_i(t) \neq 0$, we have

$$p_3(t) - p_0(\gamma(t)) = -\lambda_0(\gamma(t))\dot{p}_0(\gamma(t)), \tag{29}$$

$$p_0(t) - p_1(t) = -\lambda_1(t)\dot{p}_1(t), \tag{30}$$

$$p_1(t) - p_2(t) = -\lambda_2(t)\dot{p}_2(t), \tag{31}$$

$$p_2(t) - p_3(t) = -\lambda_3(t)\dot{p}_3(t) \tag{32}$$

for scalar functions λ_i. We have not yet shown that p_1 and p_3 are regular; however by Lemma 4, *p_0 and p_2 are regular, and* (29) *and* (31) *are valid.* Using the fact that $p_1(t)$ lies on a characteristic tangent to the convex curve p_2 and has direction orthogonal to $\dot{v}_1(t)$, it follows that $\lambda_2(t) > 0$. From (31) we find that $\|\dot{p}_2(t)\, \dot{p}_1(t)\| = -\lambda_2(t)\|\dot{p}_2(t)\, \ddot{p}_2(t)\| \neq 0$; hence $\dot{p}_1(t) \neq 0$ and p_1 is regular. Since $\dot{p}_1(t) \perp \dot{v}_1(t)$ and v_1 is convex, p_1 is convex, and (v_1, p_1) is a dual pair. Moreover, $\lambda_1(t) > 0$ and p_1 is in $\alpha_1 + \overline{\Upsilon}_1$; it follows that p_1 is in the inside of p_0. Similar results hold for p_3. We summarize these results in the following lemma.

LEMMA 5. *p_i is regular and in $\alpha_i + \overline{\Upsilon}_i$. The functions λ_i are positive. (v_i, p_i) are dual pairs. p_i is in the inside of p_{i-1}.*

Let the functions $H_i : I \to \mathbb{R}$ be dual to S_i (Definition 3). Then curves $P_i := (p_i, H_i)$ are dual to $V_i = (v_i, S_i)$ and generate tangent surfaces $\mathcal{H}_i$ dual to $\mathcal{S}_i$.

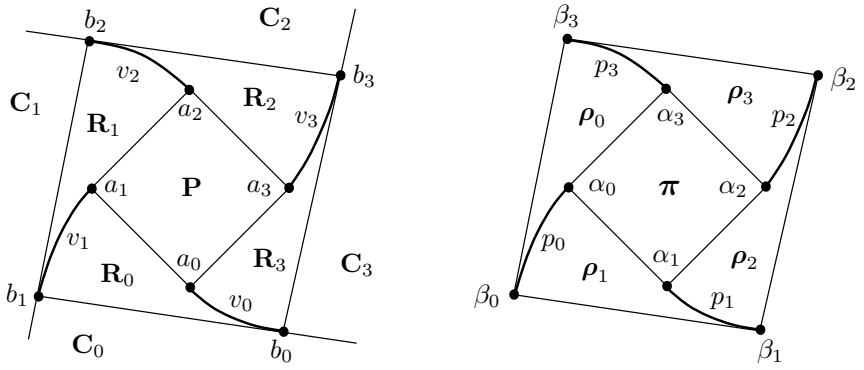


FIG. 9. *Constructed solution $\mathcal{S}$ in the v- and p-planes. Left: domains $\mathbf{C}_i$ of plane waves, $\mathbf{R}_i$ of sonic rarefactions, and $\mathbf{P}$ of a parabolic wave. $\mathcal{S}$ forms a convex edge along v_i for $i = 0, 2$ and a concave edge for $i = 1, 3$. Right: $\mathcal{H}$ is defined by $\mathcal{O}$ on $\boldsymbol{\pi}$ and curves p_i, and extended to the plane.*

PROPOSITION 7. $\mathcal{H}_{i+1}(p_i(t)) = \mathcal{H}_i(p_i(t))$.

Proof. Let $f(t) = \mathcal{H}_{i+1}(p_i(t)) - \mathcal{H}_i(p_i(t))$. From Proposition 2(b), $\dot{f}(t) = (v_{i+1}(t) - v_i(t)) \cdot \dot{p}_i(t)$ for $i = 0, 1, 2$ and $\dot{f}(t) = (v_0(\gamma(t)) - v_3(t)) \cdot \dot{p}_3(t)$ for $i = 3$. By (22)–(25), $\dot{f}(t) = 0$; hence f is a constant. Since H_i is dual to S_i, $f(0) = H_i(0) - H_{i-1}(0) - a_i \cdot (\alpha_i - \alpha_{i-1}) = 0$, thus $f \equiv 0$. ∎

Since $\lambda_i(t) > 0$, $p_{i-1}(t)$ is on the dual characteristic tangent at $p_i(t)$ and we truncate dual characteristics tangent to p_i along p_{i-1}. The union of these dual characteristics form closed sets $\boldsymbol{\rho}_i$ defined by $\boldsymbol{\rho}_i = \bigcup_{t \in I} \bigcup_{0 \le \sigma \le \lambda_i(t)} (p_i(t) - \sigma \dot{p}_i(t))$. As shown for the sets $\mathbf{R}_i$, the interiors of $\boldsymbol{\rho}_i$ and $\boldsymbol{\rho}_j$, $i \neq j$, are disjoint; hence their union is a topological annulus. We define the *dual periodic structure* $\mathcal{O}$ by $\mathcal{O}(p) = \mathcal{H}_i(p)$ for $p \in \boldsymbol{\rho}_i$. By Proposition 2(c), where it is differentiable, $\mathcal{S}$ satisfies (1) with $\mathcal{H} \equiv \mathcal{O}$.

THEOREM 5. *The dual periodic structure $\mathcal{O}$, up to a reordering of indices, is a periodic structure. The chord of $C^{\wedge\vee}(v_i(t))$ is the graph of $\mathcal{O}$ restricted to the line segment $p_{i-1}(t)p_i(t)$.*

Proof. By Proposition 7 $\mathcal{O}$ is continuous, and by Lemma 5 the curves P_i are regular, thus dual characteristics from P_i to P_{i-1} form orientation-preserving diffeomorphisms. Therefore $\mathcal{O}$ is a periodic structure with indices reversed. $C^{\wedge\vee}(v_i(t))$ is the convex hull of $p_{i-1}(t)$ and $p_i(t)$; hence the graph of $\mathcal{O}$ between these points is its chord. ∎

7. Proof of the main theorem. We now construct a function $\mathcal{S}$, Riemann data $\{\beta_i\}$, and a Hamiltonian $\mathcal{H}$ such that $\mathcal{S}$ is the Riemann solution to (1), (2). Let $\{a_i\}$, $\{\alpha_i\}$ be the dual quadrilaterals (Corollary 1) of the example of section 3. By section 6 we obtain a periodic structure composed of functions $\mathcal{S}_i$, and its dual periodic structure composed of functions $\mathcal{H}_i$, having inner orbit given by $\{a_i\}$, $\{\alpha_i\}$ and outer orbit given by $\{b_i\}$, $\{\beta_i\}$; see Figure 9. Let $\mathbf{P}$ be the convex hull of $\{a_i\}$. $\mathcal{P}$ defined by $\mathcal{P}(v) = (x^2 - y^2)/2$, $v = (x, y) \in \mathbf{P}$ is the parabolic wave of section 3. We define sets $\mathbf{C}_i = b_{i+1} + \bigcup_{\kappa, \kappa' \ge 0} \kappa(b_i - b_{i+1}) + \kappa'(b_{i+1} - b_{i+2})$, and plane waves $\mathcal{C}_i(v) = \beta_i \cdot (v - b_{i+1}) + \mathcal{S}_{i+1}(b_{i+1}), v \in \mathbf{C}_i$. $\mathbf{P}$, $\mathbf{R}_i$, and $\mathbf{C}_i$ have disjoint interiors and the solution $\mathcal{S}$ is defined by

$$\mathcal{S}(v) = \begin{cases} \mathcal{P}(v), & v \in \mathbf{P}, \\ \mathcal{S}_i(v), & v \in \mathbf{R}_i, \\ \mathcal{C}_i(v), & v \in \mathbf{C}_i. \end{cases} \tag{33}$$

This yields Riemann data $\{\beta_i\}$ by letting $v \to \infty$ in $\mathbf{C}_i$.

Let $\boldsymbol{\pi}$ be the convex hull of $\{\alpha_i\}$ and define $\boldsymbol{\pi}_0$ to be the union of $\boldsymbol{\pi}$ and the curves p_i.

PROPOSITION 8. *$\mathcal{S}$ is continuous, and at points of differentiability satisfies* (1) *with*

$$\mathcal{H}(p) = \begin{cases} (\xi^2 - \eta^2)/2, & p \text{ in } \boldsymbol{\pi}, \\ \mathcal{H}_i(p), & p \text{ on } p_i. \end{cases} \tag{34}$$

Proof. The inner orbit is the boundary of the parabolic wave and the outer orbit coincides with the plane waves, hence $\mathcal{S}$ is continuous. By Propositions 1 and 2(c), $\mathcal{S}$ satisfies (1) in $\mathbf{P}$ and $\mathbf{R}_i$, and by construction, satisfies it in $\mathbf{C}_i$. □

PROPOSITION 9. *For the given choice of dual quadrilaterals,*

$$\text{for } i = 0, 2 : \mathcal{S} \text{ satisfies Oleĭnik condition along } v_i \Leftrightarrow \mathcal{H}(p) \geq \mathcal{O}(p),\ p \in \boldsymbol{\rho}_i, \tag{35}$$

$$\text{for } i = 1, 3 : \mathcal{S} \text{ satisfies Oleĭnik condition along } v_i \Leftrightarrow \mathcal{H}(p) \leq \mathcal{O}(p),\ p \in \boldsymbol{\rho}_i. \tag{36}$$

Proof. For the choice of dual quadrilaterals, $V_i = (v_i, \mathcal{S}(v_i))$ is a convex edge for $i = 0, 2$, and is a concave edge for $i = 1, 3$. Thus $C^{\wedge\vee}(v_i(t)) = C^{\vee}(v_i(t))$ for $i = 0, 2$ and $C^{\wedge\vee}(v_i(t)) = C^{\wedge}(v_i(t))$ for $i = 1, 3$. By Theorem 5 the chord of $C^{\wedge\vee}(v_i(t))$ is given by $\mathcal{O}$ on $\boldsymbol{\rho}_i$, and the result follows (see Oleĭnik, section 2). □

Note that if we extend $\mathcal{H}$ by defining $\mathcal{H}(p) = \mathcal{O}(p)$ for $p \in \boldsymbol{\rho}_i$, and then extend further to the plane in any continuous manner, we obtain a solution with a period 4 sonic sequence for a C^0 Hamiltonian. To prove Theorem 1 we extend $\mathcal{H}$, given by (34) on $\boldsymbol{\pi}_0$, to a C^1 function such that (35), (36) hold. Equation (34) specifies derivatives of $\mathcal{H}$ on $\boldsymbol{\pi}$ and tangential derivatives on p_i. To obtain a C^1 extension we specify transverse derivatives on p_i by letting

$$D_p \mathcal{H}(p_i(t)) = v_{i+1}(t). \tag{37}$$

LEMMA 6. *$\mathcal{H}$ given by* (34), (37) *is C^1 on the closed set $\boldsymbol{\pi}_0$.*

Proof. From (34), by Proposition 2(b) and (29)–(32), $\dot{\mathcal{H}}(p_i(t)) = \dot{\mathcal{H}}_i(p_i(t)) = v_i(t) \cdot \dot{p}_i(t) = v_{i+1}(t) \cdot \dot{p}_i(t)$, agreeing with (37) and $\mathcal{H}$ is well defined on $\boldsymbol{\pi}_0$. Since $v_{i+1}(0) = a_{i+1} = D_p \mathcal{H}(\alpha_i)$ (evaluated in $\boldsymbol{\pi}$), $\mathcal{H}$ is C^1 at $p = \alpha_i$, hence C^1 on $\boldsymbol{\pi}_0$. □

For the choice (37), characteristics from v_i are directed toward v_{i+1} (see (6)) and the necessary Lax condition is satisfied. The derivatives of $\mathcal{H}$ in directions normal to the boundary of $\boldsymbol{\rho}_i$ are consistent with the Oleĭnik conditions (35), (36). Using the approximation theory of Whitney [27] there exists a C^1 extension $\mathcal{H} : \mathbb{R}^2 \to \mathbb{R}$ so that (35), (36) are satisfied.

PROPOSITION 10. *$\mathcal{S}$ is a Riemann solution to* (1) *with data $\{\beta_i\}$ and Hamiltonian $\mathcal{H} : \mathbb{R}^2 \to \mathbb{R}$.*

Proof. At infinity, $\mathcal{S}$ has gradient matching the Riemann data. By Proposition 8, $\mathcal{S}$ is continuous and a solution of (1). On the curves v_i, the Oleĭnik condition is satisfied by (35) and (36). On the boundary of $\mathbf{P}$ the one-sided gradients of $\mathcal{S}$ are the same as those in Proposition 1 and therefore the Oleĭnik condition is satisfied for $\mathcal{S}$. □

Thus the periodic structure in $\mathcal{S}$ is a period 4 sonic sequence and this proves Theorem 1.

8. Concluding remarks. We have shown that there exist C^1 Hamiltonians and Riemann data such that the corresponding Riemann problem has Riemann solution containing a period 4 sonic sequence. Our construction allows for return maps satisfying $\gamma(t) \leq t$, corresponding to inward spiraling periodic structures, and $\gamma(t) \geq t$,

corresponding to outward spiraling structures. Although the periodic sequence inner orbit was specified by square dual quadrilaterals, examination of the principal equations shows that they are well defined for arbitrary dual quadrilaterals. Thus, by using the dual quadrilaterals of the outer orbit of one periodic sequence as the inner orbit of another, the restrictions $\gamma(t) \geq t$ and $\gamma(t) \leq t$ can be relaxed, resulting in solutions with richer interval dynamics.

The C^1 Hamiltonian $\mathcal{H}$ used for the proof of Theorem 1 may be chosen to be C^∞ at points other than the four gradients α_i corresponding to the inner orbit. At these points, a convex curve joins to a straight line and thus $\mathcal{H}$ fails to be C^2. However, numerical work (discussed in [22]) suggests that solutions containing periodic sonic sequences exist for real analytic Hamiltonians, and that to prove Theorem 1 for such Hamiltonians we should consider a complex interaction between sonic rarefactions and the parabolic wave in which the parabolic wave boundary is not a straight line and characteristics leave it tangentially.

REFERENCES

[1] R. Abgrall, *Numerical discretization of the first-order Hamilton–Jacobi equation on triangular meshes*, Comm. Pure Appl. Math., 49 (1996), pp. 1339–1373.

[2] M. Bardi and L. C. Evans, *Hopf's formulas for solutions of Hamilton–Jacobi equations*, Nonlinear Anal., 8 (1984), pp. 1373–1381.

[3] M. Bardi and S. Osher, *The nonconvex multidimensional Riemann problem for Hamilton–Jacobi equations*, SIAM J. Math. Anal., 22 (1991), pp. 344–351.

[4] M. G. Crandall, L. C. Evans, and P.-L. Lions, *Some properties of viscosity solutions of Hamilton–Jacobi equations*, Trans. Amer. Math. Soc., 282 (1984), pp. 487–502.

[5] M. G. Crandall and P.-L. Lions, *Viscosity solutions of Hamilton–Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1–42.

[6] R. D. Driver, *Ordinary and Delay Differential Equations*, Springer-Verlag, New York, 1977.

[7] J. Glimm, M. J. Graham, J. Grove, X.-L. Li, T. M. Smith, D. Tan, F. Tangerman, and Q. Zhang, *Front tracking in two and three dimensions*, Comput. Math. Appl., 35 (1998), pp. 1–11.

[8] J. Glimm, C. Klingenberg, O. McBryan, B. Plohr, D. Sharp, and S. Yaniv, *Front tracking and two-dimensional Riemann problems*, Adv. Appl. Math., 6 (1985), pp. 259–290.

[9] J. Glimm, H. Kranzer, D. Tan, and F. Tangerman, *Wave fronts for Hamilton–Jacobi equations: The general theory for Riemann solutions in* R^n, Comm. Math. Phys., 187 (1997), pp. 647–677.

[10] J. Glimm, S. R. Simanca, D. Tan, F. M. Tangerman, and G. Vanderwoude, *Front tracking simulations of ion deposition and resputtering*, SIAM J. Sci. Comput., 20 (1999), pp. 1905–1920.

[11] J. Hale, *Theory of Functional Differential Equations*, Springer-Verlag, New York, 1977.

[12] S. Hamaguchi, M. Dalvie, R. T. Farouki, and S. Sethuraman, *A shock-tracking algorithm for surface evolution under reactive-ion etching*, J. Appl. Phys, 74 (1993), pp. 5172–5184.

[13] E. Hopf, *Generalized solutions of non-linear equations of first order*, J. Math. Mech., 14 (1965), pp. 951–973.

[14] H. Ishii, *Perron's method for Hamilton–Jacobi equations*, Duke Math. J., 55 (1987), pp. 369–384.

[15] F. John, *Partial Differential Equations*, Springer-Verlag, New York, 1971.

[16] P. Lax, *Hyperbolic Systems of Conservation Laws and the Mathematical Theory of Shock Waves*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 11, SIAM, Philadelphia, 1973.

[17] W. B. Lindquist, *Construction of solutions for two dimensional Riemann problems*, Internat. J. Comput. Appl. Math., 12 (1986), pp. 615–630.

[18] P.-L. Lions and J. C. Rochet, *Hopf formula and multitime Hamilton–Jacobi equations*, Proc. Amer. Math. Soc., 96 (1986), pp. 79–84.

[19] P.-L. Lions and P. E. Souganidis, *Convergence of MUSCL and filtered schemes for scalar conservation laws and Hamilton–Jacobi equations*, Numer. Math, 69 (1995), pp. 441–470.

[20] O. Oleĭnik, *On the uniqueness of the generalized solution of the Cauchy problem for a nonlinear system of equations occuring in mechanics*, Uspehi Mat. Nauk (N.S.), 12 (1957), pp. 169–176 (in Russian).

[21] S. Osher and J. Sethian, *Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton–Jacobi equations*, J. Comput. Phys, 79 (1988), pp. 12–49.
[22] J. D. Pinezich, *Numerical investigation of a Riemann problem for a Hamilton–Jacobi equation*, Experiment Math., submitted.
[23] J. D. Pinezich, *Periodic Structure in Two-Dimensional Riemann Problems for Hamilton–Jacobi Equations*, Ph.D. thesis, SUNY at Stony Brook, Stony Brook, NY, 1998.
[24] D. S. Ross, *Ion etching: An application of the mathematical theory of hyperbolic conservation laws*, J. Electrochem. Soc.: Solid-State Science and Technology, 135 (1988), pp. 1235–1240.
[25] J. A. Sethian and D. Adalsteinsson, *An overview of level set methods for etching, deposition, and lithography development*, IEEE Trans. on Semiconductor Manufacturing, 10 (1997), pp. 167–184.
[26] D. Wagner, *The Riemann problem in two space dimensions for a single conservation law*, SIAM J. Math. Anal., 14 (1983), pp. 534–559.
[27] H. Whitney, *Analytic extensions of differentiable functions defined in closed sets*, Trans. Amer. Math. Soc., 36 (1934), pp. 63–89.
[28] T. Zhang and Y. Zheng, *Two-dimensional Riemann problem for a single conservation law*, Trans. Amer. Math. Soc., 312 (1989), pp. 589–619.

SIAM J. MATH. ANAL.
Vol. 32, No. 4, pp. 707–733

EXACT MULTIPLICITY AND S-SHAPED BIFURCATION CURVE FOR SOME SEMILINEAR ELLIPTIC PROBLEMS FROM COMBUSTION THEORY*

YIHONG DU†

Abstract. In this paper, by making use of a new limiting equation and a continuation method based on a local bifurcation theorem of Crandall and Rabinowitz, we rigorously confirm some long-standing conjectures on the exact number of positive solutions for a class of elliptic equations arising from combustion theory. This work extends that of [S.-H. Wang, *Proc. Roy. Soc. London Sect. A,* 454 (1998), pp. 1031–1048] for the 1 dimension case to cover both dimensions 1 and 2, and it extends the work of [Y. Du and Y. Lou, *J. Differential Equations,* to appear] for the special case $m = 0$ to $0 \le m < 1$. It is shown that the main results are not true if the dimension is greater than 2 or if $m \ge 1$. Therefore, our results are in a sense the best possible.

Key words. perturbed Gelfand equation, bifurcation, combustion

AMS subject classifications. 35J65, 45P99, 80A25

PII. S0036141098343586

1. Introduction.

We are interested in the exact structure of the positive solution set $\{(\mu, v)\}$ of the problem

$$-\Delta v = \mu(1+\epsilon v)^m e^{v/(1+\epsilon v)} \text{ in } B,\ v = 0 \text{ on } \partial B, \tag{1.1}$$

where B is the unit open ball in R^n $(n \ge 1)$, $\epsilon > 0, m \ge 0$, are constants, and $\mu > 0$ is treated as a bifurcation parameter. By a positive solution, we mean $v > 0$ in B.

Problem (1.1) arises in combustion theory, where μ is known as the Frank–Kamenetskii parameter, v the dimensionless temperature, and ϵ the reciprocal activation energy. We refer to [4], [31], [33], and the references therein for more background.

The case that ϵ is small and $0 \le m < 1$, especially $m = 0$ (Arrhenius reaction rate) and $m = 1/2$ (bimolecular reaction rate), is of particular interest in applications and has attracted considerable amount of studies (see, e.g., [3], [4], [9], [22], [29], [31], [32], [33], and the references therein). On the basis of numerical studies, it has long been believed that the positive solution set $\{(\mu, v)\}$ of (1.1) is an S-shaped smooth curve provided that $0 \le m < 1$, ϵ is small and the space dimension $n = 1$ or 2. For dimension $n = 1$, this was rigorously proved recently by Wang [33], and for dimension $n = 2$, this was proved only for $m = 0$ by Du and Lou [9]. In this paper, among other things, we extend the result of [9] for $m = 0$ to all $0 \le m < 1$ and determine exactly how the solution curve changes once $m \ge 1$ for both dimensions 1 and 2. We also discuss briefly the higher dimension case. Our techniques can be applied to problems with more general nonlinearities, but we restrict ourselves to (1.1) for simplicity and its clear physical significance.

To explain the method we use in this paper, we must mention a closely related problem. In catalysis theory, there is an equation closely related to (1.1); under some

*Received by the editors August 14, 1999; accepted for publication (in revised form) August 22, 2000; published electronically November 10, 2000. This research was partially supported by the Australian Research Council.

http://www.siam.org/journals/sima/32-4/34358.html

†School of Mathematical and Computer Sciences, University of New England, Armidale, NSW 2351, Australia (ydu@turing.une.edu.au).

simple changes of variables, it reduces essentially to

$$-\Delta v = \mu(1-\epsilon v)^p e^{v/(1+\epsilon v)} \text{ in } B, \quad v = 0 \text{ on } \partial B, \tag{1.2}$$

where p is a nonnegative integer (see, e.g., Aris [2, Vol. 1, Chapter 4]). Let us note that when $p = 0$ and $m = 0$, both (1.1) and (1.2) reduce to the so-called perturbed Gelfand equation (see [3] for more details)

$$-\Delta v = \mu e^{v/(1+\epsilon v)};$$

and when $\epsilon = 0$, both (1.1) and (1.2) reduce to the well-known Gelfand equation

$$-\Delta v = \mu e^v.$$

For (1.2), it has been conjectured that for any nonnegative integer p, the positive solution set $\{(\mu, v)\}$ is S-shaped provided $\epsilon > 0$ is small and the dimension $n = 1$ or 2. The conjecture was proved to be true for $n = 1$ by Hastings and McLeod [17] (see also [32], [34], and [22] for further results). For $n = 2$, the conjecture is only rigorously proved by the above-mentioned work of Du and Lou [9] for $p = 0$.

Dancer [7] introduced a useful abstract perturbation method and used it to study (1.2) for small $\epsilon > 0$ and all dimension $n \geq 1$, taking the point of view that (1.2) is a perturbation of the Gelfand equation. Since the positive solution curve of the Gelfand equation on a ball is completely understood (see [19]), Dancer was able to show, among other things, that when $3 \leq n \leq 9$, (1.2) can have a large number of positive solutions at some particular values of μ if ϵ is sufficiently small. Dancer's method can be extended to (1.1). Therefore, the solution set of (1.1) in dimensions 3–9 is more complicated than S-shaped for small ϵ. In [20] and [21], asymptotic methods were used to study (1.2) with $p = 1$ and $1 \leq n \leq 3$. The numerical pictures of the bifurcation curves in these papers strongly support what was conjectured about (1.2) and agree very well with the results in [7].

Unfortunately, as we will explain below, our techniques in this paper work for (1.1) but do not seem to work for (1.2). One of the new ingredients in our approach is a new limiting problem. In order to see how this limiting problem arises naturally, as in [9], we take a different point of view to [7]. Let

$$u = \epsilon^2 v, \quad \lambda = (\epsilon^2 e^{1/\epsilon})\mu.$$

Then (1.1) becomes

$$-\Delta u = \lambda(u+\epsilon)^m e^{-1/(u+\epsilon)} \text{ in } B, \quad u = 0 \text{ on } \partial B. \tag{1.3}$$

We will regard (1.3) as a perturbation of

$$-\Delta u = \lambda u^m e^{-1/u} \text{ in } B, \quad u = 0 \text{ on } \partial B. \tag{1.4}$$

Clearly, for fixed $\epsilon > 0$, the shape of the positive solution curve $\{(\mu, v)\}$ of (1.1) is the same as that of the positive solution curve $\{(\lambda, u)\}$ of (1.3). From now on, we will focus on (1.3) rather than (1.1).

It turns out that the solution set $\{(\lambda, u)\}$ of (1.3) is determined completely by that of (1.4) when $n = 1, 2$, even for large ϵ. For $n \geq 3$, (1.4) determines part of the solution curve of (1.3). Therefore, we will have a detailed study of (1.4).

One might wonder whether such a change of variables trick can also be employed for (1.2). But unfortunately, since any positive solution v of (1.2) satisfies $\epsilon v \leq 1$, this trick does not seem promising for small ϵ.

We will actually obtain a complete understanding of the solution set of (1.4) for any $m \geq 0$ and in all dimensions. Let $n^* = (n+2)/(n-2)$ if $n > 2$ and $n^* = \infty$ if $n = 1, 2$, and let λ_1 denote the first eigenvalue of

$$-\Delta u = \lambda u, \quad u|_{\partial B} = 0.$$

Then our results on (1.4) can be summarized as follows:

- *If $m \geq n^*$ (hence $n > 2$), then* (1.4) *has no positive solution for any $\lambda > 0$.*
- *If $1 < m < n^*$, then* (1.4) *has a unique positive solution for any $\lambda > 0$.*
- *If $m = 1$, then* (1.4) *has no positive solution for $\lambda \leq \lambda_1$, and it has a unique positive solution for $\lambda > \lambda_1$.*
- *If $0 \leq m < 1$, then there exists $\lambda_0 > 0$ such that* (1.4) *has no positive solution for $\lambda < \lambda_0$; it has exactly one positive solution for $\lambda = \lambda_0$ and exactly two positive solutions for $\lambda > \lambda_0$.*

It is well known that when Ω is a ball centered at the origin, then any positive solution (λ, u) of $-\Delta u = \lambda f(u),\ u|_{\partial\Omega} = 0$ is uniquely determined by $(\lambda, u(0))$. Therefore, the positive solution curve $\{(\lambda, u)\}$ in $R \times C(\overline{B})$ is well represented by the curve $\{(\lambda, u(0))\}$ in R^2. We will also call $\{(\lambda, u(0))\}$ the positive solution curve. Moreover, we would say (λ, u) is above $(\tilde{\lambda}, \tilde{u})$ if $(\lambda, u(0))$ is above $(\tilde{\lambda}, \tilde{u}(0))$ in R^2, etc.

Our results in section 2 show that the shape of the positive solution curve $\{(\lambda, u(0))\}$ of (1.4) can be described by the diagrams in Figure 1 (see Theorems 2.5–2.7 for more details).

With the help of (1.4), we can obtain a very good understanding of (1.3) for $n = 1, 2$, while the case $n \geq 3$ is still incompletely understood. The results for $n = 1, 2$ are described by the diagrams in Figure 2 (see Theorems 3.3–3.6 for more details).

Central to our approach in this paper is a continuation method which we describe briefly below.

Set $X = C_0^{2,\alpha}(\overline{B})$, $Y = C^{\alpha}(\overline{B})$, and $F(\lambda, u) = \Delta u + \lambda f(u)$, where $f(u)$ is a smooth function. Clearly,

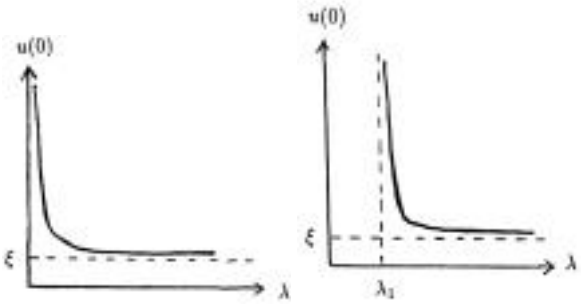
$$-\Delta u = \lambda f(u), \quad u|_{\partial B} = 0 \tag{1.5}$$

is equivalent to $F(\lambda, u) = 0$. It is easy to see that F is a smooth Fredholm mapping of index zero from $R^+ \times X$ to Y, and the partial derivative F_u at (λ, u) is given by $F_u(\lambda, u)\phi = \Delta\phi + \lambda f'(u)\phi$. If (λ_0, u_0) is not a degenerate solution, that is, if the linearization

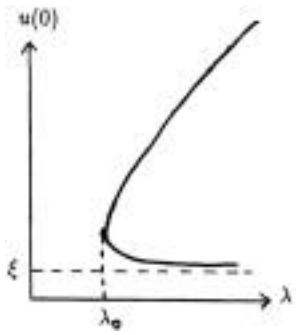
$$-\Delta\phi = \lambda_0 f'(u_0)\phi, \quad \phi|_{\partial B} = 0 \tag{1.6}$$

has no nontrivial solution ϕ, then it follows from the implicit function theorem that the solutions of (1.5) near (λ_0, u_0) form a smooth curve parametrized by λ. In other words, the solution curve can be continued from (λ_0, u_0) to both the left and the right of this point.

If (λ_0, u_0) is degenerate, i.e., (1.6) has a nontrivial solution, then the solution set of (1.5) near (λ_0, u_0) could be extremely complicated. However, if one can show that any nontrivial solution ϕ of (1.6) does not change sign in B, then the conditions of Theorem 3.2 of Crandall and Rabinowitz [5] are usually satisfied, and hence, by



(a) $1 < m < n^*$. (b) $m = 1$.



(c) $0 \le m < 1$.

FIG. 1. *Positive solution curve of* (1.4)*, where* $\xi > 0$ *when* $n > 2, \xi = 0$ *when* $n = 1, 2$.

this theorem, near the degenerate solution (λ_0, u_0), the solutions of (1.5) still form a smooth curve which is expressed in the form

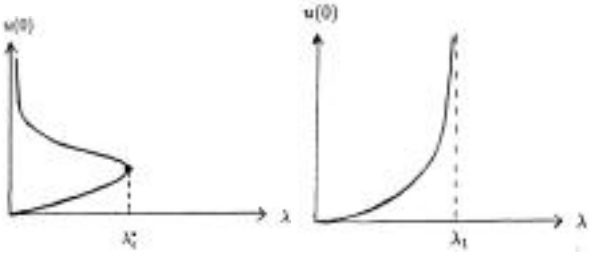
$$(\lambda(s), u(s)) = (\lambda_0 + \tau(s), u_0 + s\phi + z(s)), \tag{1.7}$$

where $s \to (\tau(s), z(s)) \in R \times Z$ is a smooth function near $s = 0$ with $\tau(0) = \tau'(0) = 0, z(0) = z'(0) = 0$, where Z is a complement of span$\{\phi\}$ in X, and ϕ is the positive solution of (1.6), which is unique if normalized.

From the expression (1.7), we see that the solution curve makes a turn to the left at (λ_0, u_0) if $\tau''(0) < 0$, and it turns to the right if $\tau''(0) > 0$. Substituting expression (1.7) to (1.5), one easily deduces that (see [1, Prop. 20.2] for a more general version of this formula)

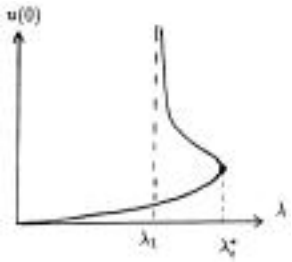
$$\tau''(0) = -\lambda_0 \frac{\int_B f''(u_0)\phi^3 dx}{\int_B f(u_0)\phi dx}. \tag{1.8}$$

Now we come to the key point of this continuation method, which allows one to determine the global shape of the bifurcation curve provided that certain information from the linearization of (1.5) can be obtained. More precisely, if one can show that whenever (1.5) has a degenerate solution, then the solutions of the linearized problem

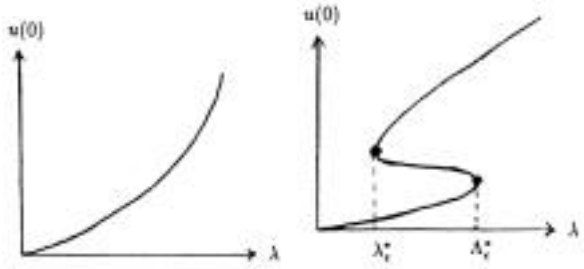


(a) $m > 1, \epsilon > 0$.
$\lim_{\epsilon \to 0} \lambda_\epsilon^* = \infty$,
$\lim_{\epsilon \to \infty} \lambda_\epsilon^* = 0$.

(b) $m = 1, \epsilon \geq 1$.



(c) $m = 1, 0 < \epsilon < 1$.
$\lim_{\epsilon \to 0} \lambda_\epsilon^* = \infty$,
$\lim_{\epsilon \to 1} \lambda_\epsilon^* = \lambda_1$.



(d) $0 \leq m < 1$,
$\epsilon \geq (1 + \sqrt{1-m})^{-2}$.

(e) $0 \leq m < 1$,
$0 < \epsilon << 1$.
$\lim_{\epsilon \to 0} \lambda_\epsilon^* = \lambda_0$,
$\lim_{\epsilon \to 0} \Lambda_\epsilon^* = \infty$.

FIG. 2. *Positive solution curve of* (1.3) *for* $n = 1, 2$.

(1.6) do not change sign, and furthermore, the corresponding $\tau''(0)$ can be proved to have the same sign by using (1.8) at any degenerate solution, then starting from any solution of (1.5), the solution curve can be continued, either by the implicit function theorem or by the Crandall–Rabinowitz theorem. More importantly, one cannot meet more than one degenerate solution in this process of continuation since the solution curve makes a turn to the same direction whenever it meets a degenerate one. Thus,

such a priori knowledge on $\tau''(0)$ determines the number of degenerate solutions on the global bifurcation curve: there is at most one degenerate solution, and hence on the whole bifurcation curve there can be at most one turning point.

The above continuation method has proved to be particularly useful in tackling bifurcation curves with exactly one turning point (see, e.g., [8], [10], [23], [28], [35], and the references therein). In section 2, we will also use this method to prove some uniqueness results. To handle S-shaped bifurcation curves in section 3, where two turning points occur on the curve, we combine this continuation method with a perturbation technique developed in [8] (see also [9], [10], and [11]).

The continuation method described above goes one step further from a well-known continuation method introduced much earlier, where the expressions (1.7) and (1.8) were used to obtain only local exact multiplicity results (see, e.g., [1], [6]). More precisely, in these earlier works, (1.7) and (1.8) were used to deal with the first turning point on the bifurcation curve which occurs when a smooth branch of stable solutions first loses stability, but the number of degenerate solutions along the global bifurcation curve is not determined, and hence the exact shape of the bifurcation curve beyond the first turning point is not determined. In this work and those mentioned in the last paragraph, the number of degenerate solutions on the bifurcation curve is determined through a priori estimates on the sign of $\tau''(0)$ at all possible degenerate solutions. This last part actually constitutes the most difficult part in the continuation method described before this paragraph and can be done only for certain nonlinear problems.

One of the referees pointed out the following recent papers, [13], [14], and [25], where the well-known continuation method is used to analyze global bifurcation branches for various nonlinear problems.

The rest of the paper is organized as follows. In section 2, we study the limiting equation (1.4). In section 3, we discuss the perturbed equation (1.3) for dimensions 1 and 2. In section 4, we discuss rather briefly the higher dimensional case for (1.3).

2. The limiting equation. In this section, we consider

$$-\Delta u = \lambda u^m e^{-1/u} \equiv \lambda f(u), \quad u|_{\partial B} = 0, \tag{2.1}$$

where $\lambda > 0,\ m \geq 0$ and $B = \{x \in R^n : |x| < 1\}, n \geq 1$. We will obtain a complete understanding for the positive solution set of (2.1). Let us note that f is C^∞ in $[0, \infty)$, including 0.

We begin with some technical but important lemmas.

LEMMA 2.1. *Suppose $0 \leq m < n^*$. If u is a degenerate positive solution of* (2.1) *and ϕ is a nontrivial solution to*

$$-\Delta\phi = \lambda f'(u)\phi, \quad \phi|_{\partial B} = 0,$$

then ϕ does not change sign in B.

The proof of Lemma 2.1 is rather long and technical. Therefore we postpone it until section 5.

LEMMA 2.2. *Suppose that u_0 is a degenerate positive solution of* (2.1) *with $\lambda = \lambda_0$. Then all positive solutions (λ, u) of* (2.1) *that are near (λ_0, u_0) in $R \times C(\overline{B})$ lie on a smooth curve represented by*

$$(\lambda, u) = (\lambda_0 + \tau(s), u_0 + s\phi + z(s)) \text{ with } s \text{ small},$$

where $z(0) = z'(0) = 0,\ \tau(0) = \tau'(0) = 0$, and ϕ is the positive eigenfunction given in Lemma 2.1. *Moreover, $\tau''(0) > 0$ if $0 \leq m < 1$, and $\tau''(0) < 0$ if $n^* > m \geq 1$.*

Proof. When $0 \le m < 1$, this follows from Lemma 2.1 and the fact that

$$f(0) = 0, \quad f''(u) > 0 \ \text{ on } (0, \alpha), \quad f''(u) < 0 \ \text{ on } (\alpha, \infty), \alpha = \frac{1}{1 - m + \sqrt{1-m}}.$$

The proof is a simple variant of [35, p. 3239] (see also [9], [23], and [28]), where the well-known formula

$$\tau''(0) = -\lambda_0 \frac{\int_B f''(u_0)\phi^3 dx}{\int_B f(u_0)\phi dx} \tag{2.2}$$

is used. We omit the details here.

When $m \ge 1$, $\tau''(0) < 0$ follows directly from (2.2) and the fact that $f''(u) > 0$ for $u > 0$. □

LEMMA 2.3. *Suppose $f \in C^1(R)$ is an arbitrary function, B is the unit ball in R^n, $n \ge 1$. Then for any given $c > 0$, the problem*

$$-\Delta u = \lambda f(u), \ u|_{\partial B} = 0$$

can have at most one solution (λ, u) satisfying $\lambda > 0, u \ge 0$ and $u(0) = c$.

Proof. This is a well-known fact. A simple proof appears in [9, Lemma 1]. □

THEOREM 2.4. *If $m \ge n^*$ and $n > 2$, then (2.1) has no positive solution for any $\lambda > 0$.*

Proof. This follows from the Pohozaev identity: If we use the notations in the proof of Lemma 2.1, then $m \ge (n+2)/(n-2)$ implies $G(u) < 0$ for $u > 0$. Thus we arrive at the contradicting inequalities before (5.3) if there is a positive solution. □

THEOREM 2.5. *Suppose $1 < m < n^*$. Then for any $\lambda > 0$, (2.1) has a unique positive solution u_λ. Moreover, $\lambda \to u_\lambda$ is a continuous (actually smooth) function from $(0, \infty)$ to $C(\overline{B})$, and $\lambda \to u_\lambda(0)$ is strictly decreasing with*

$$\lim_{\lambda \to 0^+} u_\lambda(0) = \infty, \ \lim_{\lambda \to \infty} u_\lambda(0) = \xi,$$

where $\xi > 0$ when $n > 2$ and $\xi = 0$ when $n = 1, 2$.

Proof. Clearly we have

$$\lim_{u \to 0} f(u)/u = 0, \quad \lim_{u \to \infty} f(u)/u^m = 1. \tag{2.3}$$

It follows from a standard application of the mountain pass theorem (see, e.g., [30]) that (2.1) has at least one positive solution for any $\lambda > 0$. Moreover, by the first identity in (2.3) and standard local bifurcation theory (see, e.g., [1]), any positive solution of (2.1) has its L_∞ norm bounded away from 0 for λ in any compact set of $(0, \infty)$. Due to the second identity in (2.3), the well-known blowing-up method of Gidas and Spruck [16] guarantees that any positive solution of (2.1) has its L_∞ norm bounded away from ∞ for λ in any compact set of $(0, \infty)$.

We show in the following that there is exactly one positive solution when $\lambda > 0$. Note that there is extensive literature on the uniqueness of positive solutions for various nonlinearities; see, for example, [12], [27], and the references therein. But to the best of our knowledge, our nonlinearity $f(u)$ does not seem covered by the existing results. In our proof of uniqueness below, we use a continuation argument, which seems to be new in this context.

We first prove the theorem under the assumption that any positive solution is nondegenerate. The verification of this assumption is deferred to the very end of the proof.

Under this nondegeneracy assumption, we pick up an arbitrary positive solution (λ_0, u_0) and use the implicit function theorem to continue the solution curve towards both smaller and larger values of λ and obtain a smooth solution curve $\{(\lambda, u_\lambda)\}$. As all the positive solutions are nondegenerate, this continuation procedure can be continued unless u_λ loses positivity or its L_∞ norm becomes unbounded at some finite $\lambda^* \in (0, \infty)$. However, these cases of concern cannot occur, because by Harnack inequality u_λ can lose positivity only through $\|u_\lambda\|_\infty \to 0$, which, together with the second case, is ruled out at the very beginning of the proof. Thus we obtain a smooth curve of positive solutions

$$\Gamma = \{(\lambda, u_\lambda) : 0 < \lambda < \infty\}.$$

By Lemma 2.3, the function $\lambda \to u_\lambda(0)$ must be strictly monotone. Hence

$$\lim_{\lambda \to 0^+} u_\lambda(0) = \eta \in [0, \infty].$$

From (2.3) and standard bifurcation theory we know $\eta = 0$ is impossible. If $\eta > 0$ is finite, then a simple compactness argument shows that $\lim_{\lambda \to 0^+} u_\lambda$ is a positive solution of (2.1) with $\lambda = 0$, which is evidently impossible. Therefore we must have $\eta = \infty$. It follows that $\lambda \to u_\lambda(0)$ is strictly decreasing.

Let $\xi = \lim_{\lambda \to \infty} u_\lambda(0)$. Then $\xi \in [0, \infty)$. We show that $\xi > 0$ when $n > 2$ and $\xi = 0$ when $n = 1, 2$. Indeed, when $n > 2$, $\xi > 0$ is a simple consequence of (5.3) in the proof of Lemma 2.1. When $n = 2$, we argue indirectly. Suppose that $\xi > 0$. Consider the initial value problem

$$(rz')' = -rz^m e^{-1/z}, \quad z(0) = \xi, \ z'(0) = 0.$$

It is easily seen that $z'(r) < 0$ for $r \in (0, r_0)$ as long as z is positive on $(0, r_0)$. If z remains positive on $[0, \infty)$, then $z(x) = z(|x|) = z(r)$ satisfies $\Delta z = -z^m e^{-1/z} < 0$ on R^2 and hence is a bounded subharmonic function on R^2. It is well known that in such a case, $z \equiv$ *constant*. Clearly this is impossible. Hence z has a first zero $r_0 > 0$: $z(r) > 0$ in $[0, r_0)$ and $z(r_0) = 0$. By continuous dependence of the solutions on the initial values, for λ^* large, the unique solution z^* of the initial value problem

$$(rz')' = -rz^m e^{-1/z}, \quad z(0) = u_{\lambda^*}(0), \ z'(0) = 0$$

has a first zero r^* close to r_0. But then $u^*(r) = z^*(r^* r)$ is a solution of (2.1) with $u^*(0) = u_{\lambda^*}(0)$ but $\lambda = (r^*)^2 \to r_0^2 \neq \lambda^*$ as $\lambda^* \to \infty$. This contradicts Lemma 2.3. Hence we must have $\xi = 0$.

When $n = 1$, the proof is similar but simpler. The initial value problem now is changed to

$$z'' = -z^m e^{-1/z}, \ z(0) = \xi, z'(0) = 0,$$

and the existence of a first zero of z follows from $z'' < 0$ on $[0, \infty)$.

We still need to show that Γ contains all the solutions. Suppose there is a positive solution (λ^0, u^0) not lying on Γ. Then we can repeat the above continuation argument to obtain a second solution curve $\{(\lambda, \tilde{u}_\lambda)\}$ containing (λ^0, u^0) with the property $\tilde{u}_\lambda(0) \to \infty$ as $\lambda \to 0^+$. This, however, clearly contradicts Lemma 2.3.

Finally, we verify that all the positive solutions are nondegenerate. Suppose that (λ_0, u_0) is a degenerate positive solution. It follows from Lemma 2.2 that the positive solutions near this point form a smooth curve which has a turning point at (λ_0, u_0), and the curve lies to the left of this point, i.e., all points (λ, u) on the curve have their λ values smaller than or equal to λ_0. We see that (λ_0, u_0) divides the curve into two branches. We call the branch with larger values of $u(0)$ the upper branch and denote it as $\{(\lambda, u^\lambda)\}$, while the other branch will be called the lower branch and denoted by $\{(\lambda, u_\lambda)\}$. These two branches of solutions can be continued towards smaller values of λ until we reach $\lambda = 0$ if we don't meet a degenerate solution, since, as before, this continuation cannot be stopped by loss of positivity for u or the norm of u becoming unbounded. We cannot meet a degenerate solution either, however, because by Lemma 2.2, if we meet a degenerate solution in the procedure of continuation towards smaller values of λ, then the solutions near the degenerate solution must all lie to the left of this point, which is clearly impossible.

Now we look at the lower branch

$$\{(\lambda, u_\lambda) : 0 < \lambda < \lambda_0\}.$$

By Lemma 2.3, we must have $\lambda \to u_\lambda(0)$ increasing. Therefore

$$\lim_{\lambda \to 0^+} u_\lambda(0) = \zeta \in [0, u_0(0)).$$

By (2.3) and standard bifurcation argument, we know as before $\zeta > 0$. Then a compactness argument shows $\lim_{\lambda \to 0^+} u_\lambda$ is a positive solution of (2.1) with $\lambda = 0$, a contradiction. This finishes the proof of Theorem 2.5. □

THEOREM 2.6. *If $m = 1$, then (2.1) has no positive solution for $\lambda \le \lambda_1$, and it has a unique positive solution $u = u_\lambda$ when $\lambda > \lambda_1$. Moreover, $\lambda \to u_\lambda$ is a continuous (actually smooth) function from (λ_1, ∞) to $C(\overline{B})$, and $\lambda \to u_\lambda(0)$ is strictly decreasing with*

$$\lim_{\lambda \to \lambda_1 + 0} u_\lambda(0) = \infty, \quad \lim_{\lambda \to \infty} u_\lambda(0) = \xi,$$

where $\xi > 0$ when $n > 2$ and $\xi = 0$ when $n = 1, 2$.

Proof. This is a modification of the proof for Theorem 2.5. Therefore we only point out the differences.

Since $m = 1$,

$$\lim_{u \to 0} f(u)/u = 0, \quad \lim_{u \to \infty} f(u)/u = 1. \tag{2.4}$$

It follows from Theorem 15.1 of [1] that (2.1) has at least one positive solution for $\lambda \in (\lambda_1, \infty)$. Moreover, using standard theories on local bifurcation and on asymptotically linear operators (see, e.g., [1]), one easily deduces from (2.4) that the L_∞ norm of the positive solutions are bounded away from 0 and ∞ for λ in any compact set of (λ_1, ∞).

Since

$$-\Delta u = \lambda f(u) < \lambda u,$$

one easily deduces $\lambda > \lambda_1$ whenever there is a positive solution.

The rest of the proof is exactly the same as that for Theorem 2.5 except that we replace $\lambda = 0$ there by $\lambda = \lambda_1$ now. □

THEOREM 2.7. *Suppose $0 \le m < 1$. Then there exists $\lambda_0 > 0$ such that (2.1) has no positive solution for $\lambda < \lambda_0$, exactly one positive solution for $\lambda = \lambda_0$, and exactly two positive solutions for $\lambda > \lambda_0$. Moreover, the positive solution set $\{(\lambda, u)\}$ of (2.1) forms a "$\subset$"-shaped smooth curve in the space $R \times C(\overline{B})$. Moreover, if we denote the upper and lower branches by*

$$\{(\lambda, u^\lambda) : \lambda_0 \le \lambda < \infty\} \text{ and } \{(\lambda, u_\lambda) : \lambda_0 \le \lambda < \infty\},$$

respectively, then $\lambda \to u^\lambda(x)$ is strictly increasing for any fixed $|x| < 1$, $\lambda \to u_\lambda(0)$ is strictly decreasing, and

$$\lim_{\lambda\to\infty} u^\lambda(x) = \infty \ \forall |x| < 1; \quad \lim_{\lambda\to\infty} u_\lambda(0) = \xi, \ \xi = 0 \text{ if } n = 1, 2 \text{ and } \xi > 0 \text{ if } n > 2.$$

Proof. We first show that for λ large, (2.1) has at least one positive solution. Choose $\psi \in C_0^\infty(B)$ satisfying $\psi \ge 0$ and $\max_B \psi = 1$. Let $\underline{u}$ be the unique solution of $\Delta u + \psi = 0, u|_{\partial B} = 0$, and for $\lambda > 0$ let $\overline{u}_\lambda$ be the unique positive solution of $\Delta u + \lambda u^m = 0, u|_{\partial B} = 0$. The existence here follows from a standard upper and lower solution argument since $\lim_{u\to 0} u^m/u = \infty$ and $\lim_{u\to\infty} u^m/u = 0$, while the uniqueness follows from the concavity of u^m. It is easy to check that

$$\overline{u}_\lambda = (\lambda/\lambda_0)^{1-m} \overline{u}_{\lambda_0}$$

for any positive numbers λ and λ_0. By the strong maximum principle, we see that $\overline{u}_{\lambda_0} \ge \delta \underline{u}$ for some positive constant δ. Hence $\overline{u}_\lambda \ge \underline{u}$ when λ is large. Clearly, $\Delta \overline{u}_\lambda + \lambda f(\overline{u}_\lambda) \le \Delta \overline{u}_\lambda + \lambda (\overline{u}_\lambda)^m = 0$. Thus $\overline{u}_\lambda$ is an upper solution to (2.1). On the other hand, since ψ has compact support in B while $f(\underline{u})$ is bounded away from 0 on any compact subset of the open ball B, $\lambda f(\underline{u}) \ge \psi$ on B for all large λ. It follows that $\Delta \underline{u} + \lambda f(\underline{u}) \ge \Delta \underline{u} + \psi = 0$ for large λ. Hence, for large λ, $\overline{u}_\lambda \ge \underline{u}$ and they are upper and lower solutions to (2.1), respectively. Therefore there exists $\lambda_* > 0$ such that (2.1) has at least a positive solution provided that $\lambda \ge \lambda_*$.

Now we can set

$$\lambda_0 = \inf \big\{\lambda > 0 : (2.1) \text{ has at least a positive solution}\big\}.$$

We claim that $\lambda_0 > 0$. Otherwise, there exists $\lambda_i \to 0$ and v_i positive, such that

$$\Delta v_i + \lambda_i v_i^m e^{-1/v_i} = 0, \quad v_i|_{\partial B} = 0.$$

Set $\tilde{v}_i = v_i/\|v_i\|_\infty$. Then

$$\Delta \tilde{v}_i + \lambda_i e^{-1/v_i} (v_i)^{m-1} \tilde{v}_i = 0, \quad \tilde{v}_i|_{\partial B} = 0.$$

As $e^{-1/v_i}(v_i)^{m-1}$ is uniformly bounded, by standard elliptic regularity, $\|\tilde{v}_i\|_{W^{2,p}} \to 0$. The Sobolev embedding theorem implies that $\tilde{v}_i \to 0$ uniformly. However, this is impossible as $\|\tilde{v}_i\|_\infty = 1$. This contradiction implies that $\lambda_0 > 0$.

Again by standard elliptic regularity, we can further show that (2.1) with $\lambda = \lambda_0$ has at least a positive solution. We choose one of them and denote it as u_0. We claim that u_0 must be a degenerate solution. If not, then by the implicit function theorem we can show that for λ less than but close to λ_0, (2.1) has at least one positive solution, which contradicts the definition of λ_0. Since u_0 is degenerate, our Lemma 2.2 implies that the solutions near (λ_0, u_0) form a smooth curve which turns to the right in the (λ, u) space. We may call the part of the smooth curve $\{(\lambda, u)\}$ with

$u(0) > u_0(0)$ the upper branch, and the rest the lower branch, and denote the upper and lower branches by u^λ and u_λ, respectively. As long as (λ, u^λ) and (λ, u_λ) are nondegenerate, the implicit function theorem ensures that we can continue to extend these two branches in the direction of increasing λ. To save notations, we still denote the extensions as u^λ and u_λ. This process of continuation towards larger values of λ for both branches may be stopped at some finite λ^* by one of the following three possibilities:

(i) $\|u^{\lambda_n}\|_\infty$ or $\|u_{\lambda_n}\|_\infty$ goes to infinity for some $\lambda_n \to \lambda^* - 0$;

(ii) $\|u^{\lambda_n}\|_\infty$ or $\|u_{\lambda_n}\|_\infty$ goes to 0 for some $\lambda_n \to \lambda^* - 0$ (note that by the Harnack inequality, u^λ and u_λ can only lose positivity through vanishing on the entire domain);

(iii) u^{λ^*} or u_{λ_*} is a degenerate solution.

However, (i) cannot occur since u^{λ_n} and u_{λ_n} are bounded from above by the unique positive solution of $\Delta u + (\lambda^* + 1)u^m = 0,\ u|_{\partial B} = 0$, which follows from a simple upper and lower solution argument; (ii) cannot occur either as otherwise, denoting $u_n = u^{\lambda_n}$ or u_{λ_n},

$$0 = \lambda_1(-\Delta - \lambda_n e^{-1/u_n}(u_n)^{m-1}) \to \lambda_1(-\Delta) > 0.$$

Finally, (iii) cannot occur. This is because, if, say, (λ, u^λ) becomes degenerate at $\lambda = \lambda^*$, then Lemma 2.2 tells us that all the solutions near $(\lambda^*, u^{\lambda^*})$ must lie to the right side of it, which is a contradiction. Therefore we can always extend these two branches of solutions to $\lambda = \infty$.

By Lemma 2.3, we see that the real functions $\lambda \to u^\lambda(0)$ and $\lambda \to u_\lambda(0)$ must be strictly monotone and $u^\lambda(0) > u_0(0) > u_\lambda(0)$ for any $\lambda \in (\lambda_0, \infty)$. Hence

$$\lim_{\lambda\to\infty} u_\lambda(0) = \xi \in [0, u_0(0)); \quad \lim_{\lambda\to\infty} v^\lambda(0) = \eta \in (u_0(0), \infty].$$

We show that $\eta = \infty$ always, and $\xi > 0$ when $n > 2$, $\xi = 0$ when n=1,2. By Lemma 2.3, this would imply that all the positive solutions of (2.1) are contained in these two solution branches if we can show that there is no positive solution of (2.1) satisfying $u(0) \le \xi$ when $\xi > 0$.

Let us first show that $\eta = \infty$. In fact we show a little more than that. An argument similar to but slightly simpler than that used in the proof of Lemma 3.4 in [23] shows that $\partial u^\lambda(r)/\partial\lambda > 0\ \forall r \in [0, 1)$ and $\lambda > \lambda_0$. Hence $\lambda \to u^\lambda(r)$ is strictly increasing and $u^\lambda(r) > u_0(r)$. It follows, noticing $f(u^\lambda) \ge f(u_0)$,

$$u^\lambda(r) = (-\Delta)^{-1}[\lambda f(u^\lambda)] \ge (\lambda/\lambda_0)(-\Delta)^{-1}[\lambda_0 f(u_0)] = (\lambda/\lambda_0)u_0(r) \to \infty$$

as $\lambda \to \infty$, for any $r \in [0, 1)$.

Second we show that if $\xi > 0$, then (2.1) has no positive solution with $u(0) \le \xi$. In fact, if there is such a solution, then the argument we used above can be repeated to show that there is a second smooth curve $\{(\lambda, \tilde{u})\}$ of positive solutions which is "$\subset$"-shaped and for $(\lambda, \tilde{u})$ on its upper branch, $\tilde{u}(0) \to \infty$ as $\lambda \to \infty$. This implies, however, that for any large number $C > 0$, there are at least two solutions u^λ and $\tilde{u}$ with $u^\lambda(0) = \tilde{u}(0) = C$, contradicting Lemma 2.3.

Finally, the fact that $\xi = 0$ if $n = 1, 2$ and $\xi > 0$ if $n > 2$ is proved in the same way as in the proof of Theorem 2.5.

The proof of Theorem 2.7 is now complete. □

3. The perturbed equation in dimensions 1 and 2. This section is devoted to the problem

$$-\Delta u = \lambda(u+\epsilon)^m e^{-1/(u+\epsilon)} \equiv \lambda f(u+\epsilon), \quad u|_{\partial B} = 0, \tag{3.1}$$

where $\epsilon > 0$, $m \geq 0$ and $B = \{x \in R^n : |x| < 1\}, n = 1, 2$.

Let us first observe the following simple relationship between (2.1) and (3.1).

If (λ, u) is a positive solution of (2.1), and $u(0) > \epsilon$, then we can find a unique $a \in (0,1)$ such that $u(a) = \epsilon$. Define

$$v(x) = u(ax) - \epsilon, \quad x \in B.$$

Clearly

$$-\Delta v = a^2 \lambda f(v+\epsilon), \quad v|_{\partial B} = 0.$$

That is, $(a^2\lambda, v)$ is a positive solution of (3.1).

This relationship between (2.1) and (3.1) will be frequently used in this section. Though such an exact relationship between the solutions of (2.1) and (3.1) is not essential for most of the results to be true, it simplifies the proofs substantially.

The following result will play a central role in this section.

LEMMA 3.1. *If u is a degenerate positive solution of* (3.1) *and ϕ is a nontrivial solution to*

$$-\Delta \phi = \lambda f'(u+\epsilon)\phi, \quad \phi|_{\partial B} = 0,$$

then ϕ does not change sign in B.

Proof. Before starting the proof, let us remark that our proof below requires only $\epsilon \geq 0$. Therefore, it is a simplification of the proof of Lemma 2.1 for the case $n = 1, 2$. It also simplifies the proof of Theorem 3 in [9] where only the special case $m = 0$ is considered. Let us also remark that this result is not true if $3 \leq n \leq 9$ and $\epsilon > 0$ is small (see section 4).

By [15], u is radially symmetric: $u(x) = u(r)$, $r = |x|$; moreover, $u'(r) < 0$ on (0,1]. By Proposition 3.3 of [24], ϕ is also radially symmetric: $\phi(x) = \phi(r)$. Hence

$$\phi'' + \frac{n-1}{r}\phi' + \lambda f'(u+\epsilon)\phi = 0 \text{ in } [0,1], \quad \phi'(0) = 0, \ \phi(1) = 0.$$

We may assume $\phi(0) > 0$.

As in [9], we make use of the test function

$$v(r) = ru'(r) + \mu$$

instead of the usual $v = ru' + \mu u$, where μ is a positive constant to be specified later. By a direct calculation,

$$v'' + \frac{n-1}{r}v' + \lambda f'(u+\epsilon)v = \lambda[\mu f'(u+\epsilon) - 2f(u+\epsilon)] \equiv G(r),$$

$$[r^{n-1}(v'\phi - v\phi')]' = G(r)r^{n-1}\phi, \tag{3.2}$$

where

$$G(r) = \lambda f(u+\epsilon)g(r), \quad g(r) = \mu\left[\frac{m}{u+\epsilon} + \frac{1}{(u+\epsilon)^2}\right] - 2.$$

Clearly, $g(r)$ is strictly increasing in r.

Now we suppose $\phi(r)$ changes sign in $(0,1)$ and want to deduce a contradiction from this. Let $r_0 \in (0,1)$ be the first zero of $\phi(r)$: $\phi(r_0) = 0$ and $\phi(r) > 0$ for $r \in [0, r_0)$. We choose $\mu = -r_0 u'(r_0)$ in $v = ru' + \mu$. Since

$$v' = -r\lambda f(u+\epsilon) + (2-n)u' < 0 \quad \forall r \in (0,1],$$

we have $v(r) > v(r_0) = 0$ on $[0, r_0)$ and $v(r) < 0$ on $(r_0, 1]$.

We divide our considerations below into two cases: (i) $g(r_0) \le 0$ and (ii) $g(r_0) > 0$.

In case (i), using $g(r) < g(r_0) \le 0$ on $[0, r_0)$, we obtain the following contradiction by integrating (3.2) from 0 to r_0:

$$0 > \int_0^{r_0} G(r) r^{n-1} \phi dr = [r^{n-1}(v'\phi - v\phi')]|_0^{r_0} = 0.$$

In case (ii), we consider the last zero of $\phi(r)$ before $r = 1$: $r_0 \le r^0 < 1,\ \phi(r^0) = 0, \phi(r) \ne 0$ for $r \in (r^0, 1)$. We may assume that $\phi(r) > 0$ on $(r^0, 1)$ (otherwise change the sign of ϕ). Then $\phi'(r^0) > 0 > \phi'(1)$. Now using $g(r) > 0$ and $v(r) \le 0$ on $[r^0, 1]$, we again deduce a contradiction:

$$0 < \int_{r^0}^1 G(r) r^{n-1} \phi(r) dr = [r^{n-1}(v'\phi - v\phi')]|_{r^0}^1 = (r^0)^{n-1} v(r^0)\phi'(r^0) - v(1)\phi'(1) \le 0.$$

The proof is complete. □

Using Lemma 3.1, we obtain a variant of Lemma 2.2, whose obvious proof we omit.

LEMMA 3.2. *Suppose that u_0 is a degenerate positive solution of* (3.1) *with $\lambda = \lambda_0$. Then all positive solutions (λ, u) of* (3.1) *that are near (λ_0, u_0) in $R \times C(\overline{B})$ lie on a smooth curve represented by*

$$(\lambda, u) = (\lambda_0 + \tau(s), u_0 + s\phi + z(s)) \text{ with } s \text{ small},$$

where $z(0) = z'(0) = 0,\ \tau(0) = \tau'(0) = 0$, and ϕ is the positive eigenfunction given in Lemma 3.1. *Moreover,*

$$\tau''(0) = -\lambda_0 \frac{\int_B f''(u_0+\epsilon)\phi^3 dx}{\int_B f(u_0+\epsilon)\phi dx}. \tag{3.3}$$

3.1. The case $m \ge 1$. Throughout this subsection, we assume $m \ge 1$. By Theorems 2.5 and 2.6, the solution set of (2.1) forms a smooth curve

$$\Gamma = \{(\lambda, u_\lambda) : \lambda_* < \lambda < \infty\},$$

where $\lambda_* = 0$ if $m > 1$ and $\lambda_* = \lambda_1$ if $m = 1$; moreover, $\lambda \to u_\lambda(0)$ is strictly decreasing and

$$\lim_{\lambda \to \lambda_* + 0} u_\lambda(0) = \infty, \quad \lim_{\lambda \to \infty} u_\lambda(0) = 0.$$

Therefore, given any $\epsilon > 0$, there is a unique $\lambda_\epsilon > \lambda_*$ such that

$$u_{\lambda_\epsilon}(0) = \epsilon.$$

Moreover,

$$\lim_{\epsilon\to 0}\lambda_\epsilon=\infty,\ \lim_{\epsilon\to\infty}\lambda_\epsilon\to\lambda_*,$$

and for any $\lambda\in(\lambda_*,\lambda_\epsilon)$, there is a unique $a=a_\lambda=a_\lambda(\epsilon)\in(0,1)$ such that

$$u_\lambda(a_\lambda)=\epsilon.$$

Since $u'(r)\neq 0$ on (0,1], $a_\lambda(\epsilon)$ varies smoothly with λ and ϵ. Moreover,

$$\text{for fixed } \epsilon>0,\ \lim_{\lambda\to\lambda_\epsilon-0}a_\lambda(\epsilon)=0;\quad \text{for fixed } \lambda>\lambda_*,\ \lim_{\epsilon\to 0}a_\lambda(\epsilon)=1.$$

Denote

$$v_\lambda(x)=u(a_\lambda x)-\epsilon,\ x\in B;\ \eta_\lambda=a_\lambda^2\lambda.$$

Then

$$\Gamma_\epsilon=\{(\eta_\lambda,v_\lambda):\lambda_*<\lambda<\lambda_\epsilon\}$$

is a smooth curve of positive solutions to (3.1) with

$$\lim_{\lambda\to\lambda_\epsilon-0}(\eta_\lambda,v_\lambda)\to(0,0),\ \lim_{\lambda\to\lambda_*+0}\|v_\lambda\|_\infty=\infty.$$

Hence Γ_ϵ is unbounded with one end at $(0,0)$. Since $a_\lambda\in(0,1)$, $\eta_\lambda<\lambda$ and hence Γ_ϵ lies to the left of $\lambda=\lambda_\epsilon$. Moreover, when $m>1$, it follows that

$$\eta_\lambda\to 0 \text{ as } \lambda\to 0^+.$$

If $m=1$, then it follows from $\lim_{u\to\infty}f(u+\epsilon)/u=1$ and standard analysis for asymptotically linear operators (see [1]) and $\lim_{\lambda\to\lambda_1+0}\|v_\lambda\|_\infty=\infty$ that we must have

$$\eta_\lambda\to\lambda_1 \text{ as } \lambda\to\lambda_1+0.$$

By Lemma 2.3, the fact that $\{v_\lambda(0):\lambda\in(\lambda_*,\lambda_\epsilon)\}=(0,\infty)$ implies that Γ_ϵ contains all the positive solutions of (3.1). The following result shows Γ_ϵ is exactly "$\supset$"-shaped if $m>1$.

THEOREM 3.3. *Suppose $m>1$. Then given any $\epsilon>0$, there exists $\lambda_\epsilon^*>0$ such that* (3.1) *has exactly two positive solutions for $\lambda\in(0,\lambda_\epsilon^*)$, exactly one positive solution for $\lambda=\lambda_\epsilon^*$, and no positive solution for $\lambda>\lambda_\epsilon^*$. Moreover, $\lambda_\epsilon^*\to\infty$ as $\epsilon\to 0$ and $\lambda_\epsilon^*\to 0$ as $\epsilon\to\infty$.*

Proof. Since η_λ is close to 0 when λ is close to either 0 or λ_ϵ,

$$\lambda_\epsilon^*=\max_{\lambda\in(0,\lambda_\epsilon)}\eta_\lambda$$

is achieved at some $\lambda_0\in(0,\lambda_\epsilon)$. By the implicit function theorem, $(\lambda_\epsilon^*,v_{\lambda_0})$ must be a degenerate solution of (3.1). Since $f''(u+\epsilon)>0\ \forall\ u\geq 0$, it follows from (3.3) that $\tau''(0)<0$ in Lemma 3.2 always. Thus we obtain a smooth curve of positive solutions which makes a turn to the left at $(\lambda_\epsilon^*,v_{\lambda_0})$. As before, we have an upper branch and a lower branch of the solution curve and can use the implicit function theorem to continue both branches towards smaller values of λ. Since the solutions have to go

along Γ_ϵ, the continuation procedure can only be stopped by meeting a degenerate solution. But this cannot occur because near any degenerate solution, Lemma 3.2 and the fact $\tau''(0) < 0$ imply the other solutions can only lie to the left of the degenerate one. Therefore, the upper branch continues to $(0, \infty)$ and the lower branch goes till to (0,0).

It remains to show $\lambda_\epsilon^* \to \infty$ as $\epsilon \to 0$ and $\lambda_\epsilon^* \to 0$ as $\epsilon \to \infty$. Since

$$\lambda_\epsilon^* = \eta_{\lambda_0} < \lambda_0 < \lambda_\epsilon, \text{ and } \lim_{\epsilon \to \infty} \lambda_\epsilon = 0,$$

one easily sees $\lambda_\epsilon^* \to 0$ as $\epsilon \to \infty$. On the other hand, for any fixed $\lambda > 0$,

$$\lambda_\epsilon^* \geq (a_\lambda(\epsilon))^2 \lambda \to \lambda \text{ as } \epsilon \to 0.$$

This implies $\lambda_\epsilon^* \to \infty$ as $\epsilon \to 0$. The proof is now complete. □

The case $m = 1$ is rather delicate. The following result shows the solution curve Γ_ϵ changes from "$\supset$"–shaped to a monotone curve when ϵ crosses $\epsilon = 1$ from $\epsilon < 1$.

THEOREM 3.4. *Suppose $m = 1$.*

(a) *If $\epsilon \geq 1$, then (3.1) has a unique positive solution u_λ when $\lambda \in (0, \lambda_1)$, and it has no positive solution for $\lambda \geq \lambda_1$. Moreover, $u_\lambda(0) \to \infty$ as $\lambda \to \lambda_1$.*

(b) *If $\epsilon \in (0, 1)$, then there exists $\lambda_\epsilon^* > \lambda_1$ such that (3.1) has exactly one positive solution for $\lambda \in (0, \lambda_1] \cup \{\lambda_\epsilon^*\}$, exactly two positive solutions for $\lambda \in (\lambda_1, \lambda_\epsilon^*)$, and no positive solution for $\lambda > \lambda_\epsilon^*$. Moreover, $\lambda_\epsilon^* \to \infty$ as $\epsilon \to 0$ and $\lambda_\epsilon^* \to \lambda_1$ as $\epsilon \to \infty$.*

Proof. (a) A simple calculation shows

$$[f(u+\epsilon)/u]' = e^{-1/(u+\epsilon)} \frac{u - \epsilon u - \epsilon^2}{u^2(u+\epsilon)} < 0 \quad \forall\, u > 0 \text{ if } \epsilon \geq 1.$$

By [18], this implies that (3.1) has at most one positive solution for any $\lambda > 0$. On the other hand, we have the positive solution curve Γ_ϵ connecting (0,0) and (λ_1, ∞). The conclusion of part (a) now follows from these two facts.

(b) Define

$$\lambda_\epsilon^* = \sup_{\lambda \in (\lambda_1, \lambda_\epsilon)} \eta_\lambda.$$

We show that $\lambda_\epsilon^* > \lambda_1$ when $\epsilon \in (0, 1)$. It suffices to show $\eta_\lambda > \lambda_1$ for some $\lambda \in (\lambda_1, \lambda_\epsilon)$. Choose $\lambda^n \to \lambda_1 + 0$ and denote

$$\mu_n = \eta_{\lambda^n}, \quad u_n = v_{\lambda^n}.$$

We know from the discussion before Theorem 3.3 that $\mu_n \to \lambda_1$ and $\|u_n\|_\infty \to \infty$. A simple compactness argument reveals $u_n/\|u_n\|_\infty \to \phi_1$ in C^1, where ϕ_1 is the first eigenfunction:

$$-\Delta\phi_1 = \lambda_1\phi_1, \ \phi_1|_{\partial B} = 0, \ \phi_1 > 0, \|\phi_1\|_\infty = 1.$$

Multipling $-\Delta u_n = \mu_n f(u_n + \epsilon)$ by ϕ_1, integrating over B, and using integration by parts yields

$$\lambda_1 \int_B u_n \phi_1 dx = \mu_n \int_B (u_n + \epsilon) e^{-1/(u_n+\epsilon)} \phi_1 dx.$$

Using the elementary inequality $e^{-t} \le 1 - t + t^2/2 \ \forall\ t \ge 0$, we deduce

$$\lambda_1 \int_B u_n \phi_1 dx \le \mu_n \int_B (u_n + \epsilon) \left[1 - \frac{1}{u_n + \epsilon} + \frac{1}{2(u_2 + \epsilon)^2}\right] \phi_1 dx$$

$$= \mu_n \int_B \left[u_n + \epsilon - 1 + \frac{1}{2(u_n + \epsilon)}\right] \phi_1 dx.$$

Hence, due to $u_n/\|u_n\|_\infty \to \phi_1$,

$$(\mu_n - \lambda_1) \int_B u_n \phi_1 dx \ge \mu_n \int_B \left[1 - \epsilon - \frac{1}{2(u_n + \epsilon)}\right] \phi_1 dx \to \lambda_1 (1 - \epsilon) \int_B \phi_1 dx > 0.$$

This implies that $\mu_n = \eta_{\lambda^n} > \lambda_1$ for large n. Therefore, we have proved $\lambda^*_\epsilon > \lambda_1$. Since η_λ is close to 0 for λ near λ_ϵ and it converges to λ_1 as $\lambda \to \lambda_1$, λ^*_ϵ must be achieved at some $\lambda_0 \in (\lambda_1, \lambda_\epsilon)$.

Now the argument used in the proof of Theorem 3.3 can be repeated to show that the solution curve can be continued from $(\lambda^*_\epsilon, v_{\lambda_0})$ leftwards, with the lower branch reaching (0,0) and the upper branch reaching (λ_1, ∞). Also, the fact that $\lambda^*_\epsilon \to \infty$ as $\epsilon \to 0$ follows from the same argument as that in the proof of Theorem 3.3. To show $\lambda^*_\epsilon \to \lambda_1$ as $\epsilon \to 1$, we use an indirect argument. Suppose $\lambda^*_{\epsilon_n} \to \lambda_0 > \lambda_1$ for some sequence $\epsilon_n \to 1$. Denote by u_n the positive solution of (3.1) at $\lambda = \lambda^*_{\epsilon_n}$. Then one easily sees by a compactness argument that u_n has a subsequence which converges to a positive solution of (3.1) with $\epsilon = 1$ and $\lambda = \lambda_0 > \lambda_1$. This contradicts part (a). The proof is complete. □

3.2. The case $0 \le m < 1$. In this subsection, we assume $0 \le m < 1$. By Theorem 2.7, the solution curve of (2.1) is "⊂"-shaped with exactly one turning point at (λ_0, u_0), where $u_0 = u_{\lambda_0} = u^{\lambda_0}$. Denote $\xi_0 = u_0(0)$. Then for any $\epsilon \ge \xi_0$, we can find a unique $\lambda^\epsilon \in [\lambda_0, \infty)$ such that

$$u^{\lambda^\epsilon}(0) = \epsilon.$$

By Theorem 2.7, for any $\lambda > \lambda^\epsilon$, we can find a unique $a^\lambda = a^\lambda(\epsilon) \in (0, 1)$ such that

$$u^\lambda(a^\lambda) = \epsilon.$$

Moreover, $\lambda \to a^\lambda$ is strictly increasing and

$$\lim_{\lambda \to \lambda^\epsilon - 0} a^\lambda = 0, \ \lim_{\lambda \to \infty} a^\lambda = 1.$$

As before, define

$$\eta^\lambda = (a^\lambda)^2 \lambda, \ v^\lambda(x) = u(a^\lambda x) - \epsilon, \ x \in B.$$

Then

$$\Gamma^\epsilon = \{(\eta^\lambda, v^\lambda) : \lambda^\epsilon < \lambda < \infty\}$$

gives a smooth solution curve of (3.1). Since a^λ is increasing with λ, it follows η^λ is strictly increasing with λ. Therefore, Γ^ϵ is a monotone curve connecting $(0, 0)$ (when $\lambda \to \lambda^\epsilon$) and (∞, ∞) (when $\lambda \to \infty$).

If $\epsilon \in (0, \xi_0)$, then for any $\lambda \geq \lambda_0$, we can find $a^\lambda = a^\lambda(\epsilon) \in (0,1)$ satisfying $u^\lambda(a^\lambda) = \epsilon$, and define η^λ, v^λ as above to obtain a smooth monotone solution curve of (3.1):

$$\Gamma^\epsilon = \{(\eta^\lambda, v^\lambda) : \lambda_0 \leq \lambda < \infty\}.$$

Clearly Γ^ϵ connects $(\eta^{\lambda_0}, v^{\lambda_0})$ to (∞, ∞).

Moreover, since $\epsilon < \xi_0$, we can find a unique $\lambda_\epsilon > \lambda_0$ such that

$$u_{\lambda_\epsilon}(0) = \epsilon.$$

By Theorem 2.7, we see that λ_ϵ increases as ϵ decreases and $\lambda_\epsilon \to \infty$ as $\epsilon \to 0$.

For any $\lambda \in [\lambda_0, \lambda_\epsilon)$, we can find a unique $a_\lambda = a_\lambda(\epsilon)$ such that

$$u_\lambda(a_\lambda) = \epsilon.$$

Clearly, for any fixed $\lambda \geq \lambda_0$,

$$\lim_{\epsilon \to 0} a_\lambda(\epsilon) = 1.$$

Now we define

$$\eta_\lambda = (a_\lambda)^2 \lambda, \ v_\lambda(x) = u_\lambda(a_\lambda x) - \epsilon, \ x \in B$$

and find that

$$\Gamma_\epsilon = \{(\eta_\lambda, v_\lambda) : \lambda_0 \leq \lambda < \lambda_\epsilon\}$$

gives another piece of smooth solution curve to (3.1). Moreover, Γ_ϵ connects the end point $(\eta^{\lambda_0}, v^{\lambda_0})$ of Γ^ϵ (when $\lambda = \lambda_0$) and $(0,0)$ (when $\lambda \to \lambda_\epsilon - 0$). Thus

$$\Gamma(\epsilon) = \Gamma^\epsilon \cup \Gamma_\epsilon$$

gives a piecewise smooth (in fact, smooth) curve for (3.1) connecting (0,0) and (∞, ∞). By Lemma 2.3, we know it contains all the positive solutions of (3.1). We are going to find out the shape of this curve.

A straightforward calculation gives

$$[f(u+\epsilon)/u]' = \frac{(u+\epsilon)^{m-2}}{u^2} e^{-1/(u+\epsilon)} [-(u+\epsilon)^2 + mu(u+\epsilon) + u].$$

It follows from elementary analysis that

$$[f(u+\epsilon)/u]' \leq 0 \ \text{ if } \epsilon \geq \epsilon_0 \equiv (1 + \sqrt{1-m})^{-2}.$$

Therefore, by [18], for any $\lambda > 0$, (3.1) has at most one positive solution if $\epsilon \geq \epsilon_0$. It follows that if $\xi_0 > \epsilon_0$, then $\Gamma(\epsilon)$ must be a monotone curve when $\epsilon \geq \epsilon_0$.

Summarizing the above discussions, we obtain the following result.

THEOREM 3.5. *If $0 \leq m < 1$ and $\epsilon \geq \min\{\xi_0, (1+\sqrt{1-m})^{-2}\}$, then (3.1) has a unique positive solution for any $\lambda > 0$.*

Our next result shows that $\Gamma(\epsilon)$ is exactly S-shaped if $\epsilon > 0$ is sufficiently small.

THEOREM 3.6. *Suppose $0 \leq m < 1$. Then for all sufficiently small $\epsilon > 0$, the solution curve $\Gamma(\epsilon)$ of (3.1) is exactly S-shaped: There exist λ^*_ϵ and Λ^*_ϵ satisfying*

(i) $0 < \lambda^*_\epsilon < \Lambda^*_\epsilon < \infty$;

(ii) (3.1) *has a unique positive solution for* $\lambda \in (0, \lambda^*_\epsilon) \cup (\Lambda^*_\epsilon, \infty)$*;*
(iii) (3.1) *has exactly two positive solutions for* $\lambda = \lambda^*_\epsilon$ *and* $\lambda = \Lambda^*_\epsilon$*;*
(iv) (3.1) *has exactly three positive solutions for* $\lambda \in (\lambda^*_\epsilon, \Lambda^*_\epsilon)$*;*
(v) $\lim_{\epsilon\to 0} \lambda^*_\epsilon = \lambda_0$, $\lim_{\epsilon\to 0} \Lambda^*_\epsilon = \infty$.

Proof. Recall that

$$f''(u) > 0 \text{ for } u \in (0, \alpha), \text{ where } \alpha = 1/[(1-m) + \sqrt{1-m}].$$

We fix some $\xi_1 \in (0, \alpha)$ and suppose

$$\epsilon < \epsilon_1 \equiv \alpha - \xi_1.$$

Then clearly $f''(u + \epsilon) > 0$ for $u \in (0, \xi_1)$.

Now we choose $\lambda_{\xi_1} > \lambda_0$ such that

$$u_\lambda(0) < \xi_1 \text{ when } \lambda \geq \lambda_{\xi_1},$$

where u_λ is the positive solution of (2.1) lying on the lower branch. By shrinking ϵ_1 we may assume that $\lambda_{\xi_1} < \lambda_\epsilon \ \forall\ \epsilon \in (0, \epsilon_1)$. We can now divide Γ_ϵ into two parts:

$$\Gamma^1_\epsilon = \{(\eta_\lambda, v_\lambda) : \lambda_{\xi_1} \leq \lambda < \lambda_\epsilon\};\ \Gamma^2_\epsilon = \{(\eta_\lambda, v_\lambda) : \lambda_0 \leq \lambda \leq \lambda_{\xi_1}\}.$$

We first analyze the shape of Γ^1_ϵ. Define

$$\Lambda^*_\epsilon = \sup_{\lambda\in[\lambda_{\xi_1}, \lambda_\epsilon)} \eta_\lambda.$$

One easily shows that there exists $\epsilon_2 \in (0, \epsilon_1]$ such that when $\epsilon \in (0, \epsilon_2)$,

$$\Lambda^*_\epsilon \text{ is achieved at some } \lambda_* \in (\lambda_{\xi_1}, \lambda_\epsilon) \text{ and } \lim_{\epsilon\to 0} \Lambda^*_\epsilon = \infty.$$

By the implicit function theorem, $(\eta_{\lambda_*}, v_{\lambda_*})$ must be a degenerate solution of (3.1). Then by Lemma 3.2, (3.3), and our choice of ξ_1, the solutions of (3.1) near $(\eta_{\lambda_*}, v_{\lambda_*})$ has a turn to the left. Therefore, we have an upper branch and a lower branch of positive solutions starting from this point, and both branches can be continued towards smaller values of λ. The lower branch can be continued to reach (0,0), because (a) we cannot meet a degenerate solution in the way of continuation due to Lemma 3.2 and $u(0) < \xi_1$ on Γ^1_ϵ, and (b) the branch goes along Γ^1_ϵ. For the same reason, the upper branch can be continued until it reaches $(\eta_{\lambda_{\xi_1}}, v_{\lambda_{\xi_1}})$. This implies that Γ^1_ϵ is exactly "$\supset$"-shaped.

Next we analyze the shape of Γ^2_ϵ. It is more convenient for our discussion if we consider a bigger piece of solution curve

$$\Gamma^3_\epsilon = \Gamma^2_\epsilon \cup \{(\eta^\lambda, v^\lambda) : \lambda_0 \leq \lambda \leq \lambda_{\xi_1}\},$$

which contains part of Γ^ϵ. We observe that any $(\lambda, u) \in \Gamma^3_\epsilon$ satisfies

$$0 < \lambda^*_\epsilon \leq \lambda \leq \lambda_{\xi_1},\ \ u_{\lambda_{\xi_1}}(0) - \epsilon \leq \|u\|_\infty = u(0) \leq u^{\lambda_{\xi_1}}(0) - \epsilon, \tag{3.4}$$

where

$$\lambda^*_\epsilon = \inf\{\lambda : (\lambda, u) \in \Gamma^3_\epsilon\}.$$

It is easily seen that λ_ϵ^* is achieved at some $\eta_{\lambda'}$, $\lambda' \in [\lambda_0, \lambda_{\xi_1})$. Therefore $(\lambda_\epsilon^*, v_{\lambda'})$ must be a degenerate solution of (3.1). Clearly

$$\lambda_\epsilon^* \le \eta_{\lambda_0} = (a_{\lambda_0}(\epsilon))^2 \lambda_0 < \lambda_0.$$

On the other hand, it is easy to see that $a_\lambda(\epsilon) \to 1$ as $\epsilon \to 0$ uniformly for $\lambda \in [\lambda_0, \lambda_{\xi_1}]$. Hence

$$\lim_{\epsilon \to 0} \lambda_\epsilon^* = \lim_{\epsilon \to 0} \min\{(a_\lambda(\epsilon))^2 \lambda : \lambda_0 \le \lambda \le \lambda_{\xi_1}\} = \lambda_0.$$

We know from the discussion above that Γ_ϵ^3 contains at least one degenerate solution $(\lambda_\epsilon^*, v_{\lambda'})$. If we can show that there exists $\epsilon_3 \in (0, \epsilon_2)$ such that whenever $\epsilon \in (0, \epsilon_3)$, any degenerate solution on Γ_ϵ^3 must make $\tau''(0) > 0$ in (3.3) of Lemma 3.2, then a continuation argument much as before shows Γ_ϵ^3 contains exactly one degenerate solution at $\lambda = \lambda_\epsilon^*$ and the curve makes a turn to the right at this point. Hence Γ_ϵ^3 must be "$\subset$"-shaped. This tells us that the entire solution curve $\Gamma(\epsilon)$ is exactly S-shaped with two turning points at $\lambda = \lambda_\epsilon^*$ and $\lambda = \Lambda_\epsilon^*$, respectively. Clearly, this would finish the proof of Theorem 3.6.

It remains to show that there exists $\epsilon_3 \in (0, \epsilon_2)$ such that any degenerate solution on Γ_ϵ^3 must make $\tau''(0) > 0$ in (3.3) of Lemma 3.2 as long as $\epsilon \in (0, \epsilon_3)$. We argue indirectly. Suppose for some $\epsilon_k \to 0$, we can find degenerate solutions $(\lambda^k, u^k) \in \Gamma_{\epsilon_k}^3$ such that

$$\tau_k''(0) = -\lambda^k \frac{\int_B f''(u^k + \epsilon_k)\phi_k^3 dx}{\int_B f(u^k + \epsilon)\phi_k dx} \le 0,$$

where ϕ_k is the positive eigenfunction given in Lemma 3.1 when $(\lambda, u) = (\lambda^k, u^k)$. We may assume that $\|\phi_k\|_\infty = 1$.

By (3.4), we may assume that $\lambda^k \to \lambda^0 \in [\lambda_0, \lambda_{\xi_1}]$. The second part of (3.4) implies that $\|f(u^k + \epsilon_k)\|_\infty$ is uniformly bounded. Therefore, by the equation for u^k and a standard regularity and compactness argument, $\{u^k\}$ has a convergent subsequence in C^1. We may assume $u^k \to u^0$ in C^1. Moreover, from

$$-\Delta\phi_k = \lambda^k f'(u^k + \epsilon_k)\phi_k, \ \phi_k|_{\partial B} = 0,$$

we can use a similar regularity and compactness argument to obtain a C^1 convergent subsequence of ϕ_k. We may assume $\phi_k \to \phi^0$. Then we easily deduce

$$-\Delta u^0 = \lambda^0 f(u^0), \ u^0|_{\partial B} = 0, \ u^0 \ge 0, u^0 \ne 0$$

and

$$-\Delta\phi^0 = \lambda^0 f'(u^0)\phi^0, \ \phi|_{\partial B} = 0, \phi^0 \ge 0, \|\phi^0\|_\infty = 1.$$

This is to say (λ^0, u^0) is a degenerate positive solution of (2.1) and ϕ^0 is the corresponding positive eigenfunction. By Theorem 2.7, (2.1) has a unique degenerate positive solution which is (λ_0, u_0), and by Lemma 2.2 and (2.2),

$$\tau''(0) = -\lambda_0 \frac{\int_B f''(u_0)\phi^3 dx}{\int_B f(u_0)\phi dx} > 0.$$

Therefore, we must have $\lambda^k \to \lambda_0, u^k \to u_0$ and $\phi^0 = \phi$ (note that the positive eigenfunction is unique if it is normalized). Then we deduce, however,

$$0 \ge \tau_k''(0) = -\lambda^k \frac{\int_B f''(u^k + \epsilon_k)\phi_k^3 dx}{\int_B f(u^k + \epsilon)\phi_k dx} \to -\lambda_0 \frac{\int_B f''(u_0)\phi^3 dx}{\int_B f(u_0)\phi dx} > 0.$$

This contradiction finishes our proof. □

4. Some remarks on the perturbed equation in higher dimensions. If $n \geq 3$, then the results of section 2 imply that any positive solution (λ, u) of (1.4) satisfies

$$u(0) > \xi > 0, \text{ where } \xi = \lim_{\lambda \to \infty} u_\lambda(0),$$

as in Theorems 2.5–2.7. Therefore, the trick of using positive solutions of (1.4) to obtain that for (1.3) mentioned at the beginning of section 3 can only give positive solutions (λ, u) of (1.3) with $u(0) > \xi - \epsilon$.

On the other hand, since along the positive solution curve Γ of (1.4), all the solutions are nondegenerate except at most one which, if exists, satisfies condition A of [7, p. 413] due to Lemma 2.2, we see that Theorem 2 of [7] and the results on page 412 of [7] (see also an improvement of these results in [11, Proposition A3]) can be used to conclude the following:

> *For any compact part Γ_c of Γ, there exists an $\epsilon_0 > 0$ small such that for each $\epsilon \in (0, \epsilon_0)$, there is a piece of solution curve Γ_c^ϵ of (1.3) which is close to Γ_c and has exactly the same shape as Γ_c, i.e., Γ_c^ϵ makes exactly the same number of turns as Γ_c does.*

In particular, if $0 \leq m < 1$, then Γ_c^ϵ is exactly "$\subset$"-shaped. Moreover, in this case, we can use the construction in section 3.2 to see that the upper branch of Γ_c^ϵ can be continued monotonically to (∞, ∞). Thus, when $0 \leq m < 1$ and ϵ is small, we can obtain a complete understanding of the part of positive solution curve of (1.3) where $u(0) > \xi$. Note, however, in the scale of (1.1), $v = \epsilon^{-2} u > \xi/\epsilon^2$ is large.

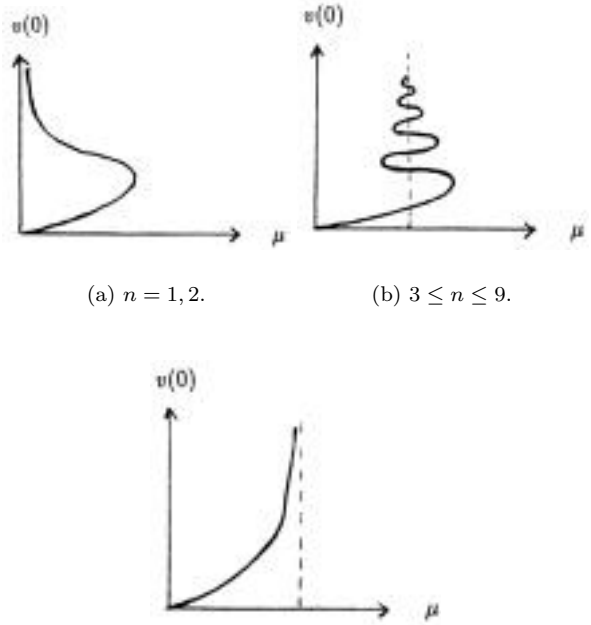
Let us now switch to (1.1) in order to understand the rest of the positive solution curve of (1.3). We now take Dancer's point of view and regard (1.1) as a perturbation of the Gelfand equation. Let Γ_0 denote the positive solution curve of the Gelfand equation on the unit ball B with Dirichlet boundary conditions. It is well known that Γ_0 can be described by the diagrams in Figure 3.

Dancer proved in [7] that solutions on Γ_0 are either nondegenerate or degenerate but satisfy his condition A. Therefore, as before, by the perturbation results in [7], for any compact part Γ_0^c of Γ_0, there exists $\epsilon_1 > 0$ small such that for each $\epsilon \in (0, \epsilon_1)$, (1.1) has a piece of positive solution curve Γ_ϵ^c which is close to Γ_0^c and makes exactly the same number of turns. In particular, in dimensions $3 \leq n \leq 9$, Γ_ϵ^c makes a large number of turns for certain μ.

Moreover, it follows from arguments in [7, p. 430] (see also Theorem 2.26 of [26]) that the eigenfunction of the linearization of the Gelfand equation at each turning point of its solution curve (except the first turning point) changes sign. Using this fact and an argument similar to that near the end of the proof of Theorem 3.6 in the present paper, one can see that the eigenfunction in Lemma 3.1 with small positive ϵ may change sign when $3 \leq n \leq 9$.

Using a phase plane argument as in [7], and employing the graph for $\{(\mu, v'(1))\}$ of the Gelfand problem, one can show as in [7] that the solution curve Γ_ϵ^c can be continued for larger values of $v(0)$ until $\mu > 0$ small (see Figure 1 in [7]), and this further continued part Γ_ϵ' has the shape similar to that of Γ_ϵ^c, but as was pointed out in [7], it is not clear if Γ_ϵ' can make extra turns besides the ones corresponding to that of Γ_ϵ^c. We illustrate these by Figure 4 (compare with Figure 1 in [21]) while omitting the rigorous proofs.

As can be seen from Figure 4, the solutions $(\mu, v) \in \Gamma_\epsilon'$ have $v'(1)$ small while that on Γ_c^ϵ obtained from (1.4) have $v'(1) = \epsilon^{-2} u_\lambda'(1)$. Therefore, there is still a gap between $\Gamma_\epsilon \cup \Gamma_\epsilon'$ and the rescaled Γ_c^ϵ.



(a) $n = 1, 2.$ (b) $3 \le n \le 9.$

(c) $n \ge 10.$

FIG. 3. *Positive solution curve of the Gelfand equation on the unit ball.*

Nevertheless, by our calculations in section 3.2, $[f(u+\epsilon)/u]' \le 0$ if $0 \le m < 1$ and $\epsilon \ge (1+\sqrt{1-m})^{-2}$. Therefore it is easy to show that (1.3) (and hence (1.1)) has a unique positive solution for any $\lambda > 0$ in this case.

5. The proof of Lemma 2.1. In this section, we give the rather long and technical proof of Lemma 2.1.

Proof of Lemma 2.1. By [15], u is radially symmetric: $u(x) = u(r)$, $r = |x|$; moreover, $u'(r) < 0$ on (0,1]. By Proposition 3.3 of [24], ϕ is also radially symmetric: $\phi(x) = \phi(r)$. Hence

$$\phi'' + \frac{n-1}{r}\phi' + \lambda f'(u)\phi = 0 \text{ in } [0,1], \quad \phi'(0) = 0, \ \phi(1) = 0.$$

By the Harnack inequality (or a well-known uniqueness result for the above singular second order ordinary differential equation), $\phi(0) \ne 0$. We may assume $\phi(0) > 0$.

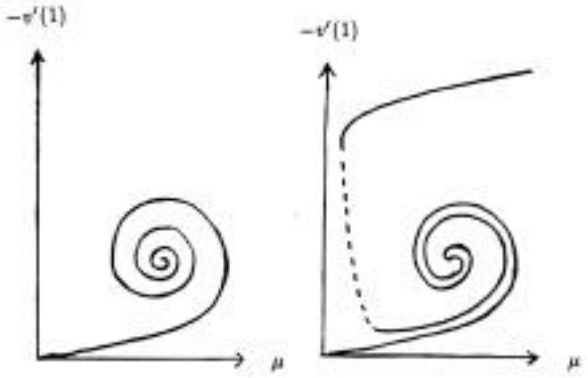
Direct calculations give

$$f'(u) = u^{m-2}e^{-1/u}(mu+1) > 0 \ \forall u > 0,$$

$$f''(u) = u^{m-4}e^{-1/u}[m(m-1)u^2 + 2(m-1)u + 1].$$

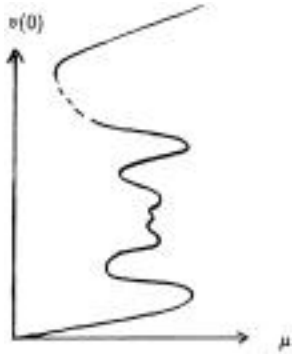
Hence

$$f''(u) > 0 \text{ for } u \in (0,\alpha); \quad f''(u) < 0 \text{ for } u \in (\alpha,\infty),$$



(a)
Gelfand equation with $3 \le n \le 9$.

(b)
(1.1) with $3 \le n \le 9$, $0 < \epsilon << 1$.



(c)
(1.1) with $3 \le n \le 9$, $0 < \epsilon << 1$.

FIG. 4.

where $\alpha = 1/[(1-m) + \sqrt{1-m}]$ if $m < 1$, and $\alpha = \infty$ if $m \ge 1$.

One easily sees

$$K(u) = uf'(u)/f(u) = m + 1/u$$

is a decreasing function of u on $(0, \infty)$ and

(a) if $m < 1$, then $K(\beta) = 1$ for $\beta = 1/(1-m) > \alpha > \rho \equiv \alpha - f(\alpha)/f'(\alpha)$;

(b) if $m \ge 1$, then $K(u) > 1 \ \forall \ u > 0$.

We define $\beta = \infty$ in case (b) and divide our discussion below into two cases:

$$\text{(i) } u(0) \le \beta \text{ and (ii) } u(0) > \beta.$$

Consider case (i) *first.* Let

$$v(r) = ru_r(r) + \mu u(r),$$

where μ is a positive constant to be specified later. Then

$$\begin{aligned} -\Delta v - \lambda f'(u)v &= \lambda[2f(u) - \mu(f'(u)u - f(u)] \\ &= \lambda f(u)[2 - \mu(K(u) - 1)]. \end{aligned} \tag{5.1}$$

Define

$$h(r) = -ru'(r)/u(r), \quad r \in [0, 1).$$

Clearly $h(0) = 0$ and $h(1) = +\infty$. We show in the following that $h'(r) > 0$ in (0,1). Indeed,

$$h'(r) = [ru_r^2 + (n-2)u_r u + \lambda r f(u)u]/u^2 = [2H(r) + \lambda r Q(u(r))]/u^2, \tag{5.2}$$

where

$$H(r) = [ru_r^2 + (n-2)u_r u]/2 + \lambda r F(u(r)), \ F(u) = \int_0^u f(s)ds,$$

$$Q(u) = uf(u) - 2F(u).$$

Here and in what follows, u_r is sometimes used for u' to avoid notations like u'^2.

If $n = 1, 2$, then it follows from the first equality in (5.2) that $h'(r) > 0$ in $(0, 1)$. Therefore, we need only consider $n > 2$ below. A simple calculation gives

$$[r^{n-1}H(r)]' = \lambda r^{n-1}G(u(r)) \text{ with } G(u) = nF(u) - \frac{n-2}{2}f(u)u.$$

Clearly $G(0) = 0$ and

$$\begin{aligned} G'(u) &= \tfrac{n+2}{2}f(u) - \tfrac{n-2}{2}f'(u)u \\ &= u^{m-1}e^{-1/u}\{[(n+2) - m(n-2)]u - (n-2)\}/2. \end{aligned}$$

It follows that

$$G'(u) < 0 \text{ on } [0, \gamma); \quad G'(u) > 0 \text{ on } (\gamma, \infty),$$

where $\gamma = (n-2)/[(n+2) - m(n-2)]$. Therefore, we have either $G(u) < 0$ on $(0, \infty)$ or $G(u) < 0$ on $(0, \gamma_0)$ and $G(u) > 0$ on (γ_0, ∞) for some $\gamma_0 > \gamma$. We show that actually only the latter alternative can occur. Indeed, if $G(u(r)) \le 0 \ \forall \ r \in [0, 1]$, then

$$0 < u_r^2(1)/2 = H(1) = \int_0^1 \lambda r^{n-1}G(u(r))dr \le 0.$$

This contradiction shows γ_0 exists and moreover

$$u(0) > \gamma_0 \tag{5.3}$$

whenever $n > 2$ and u is a positive solution of (2.1).

Let t be uniquely determined by $u(t) = \gamma_0$. We have

$$[r^{n-1}H(r)]' = \lambda r^{n-1}G(u(r)) > 0 \ \ \forall r \in (0, t),$$

which implies $H(r) > 0$ for $r \in (0,t]$. Moreover, for $r \in (t,1]$, $G(u(r)) < 0$ and therefore

$$\begin{aligned} r^{n-1}H(r) &= H(1) - \int_r^1 \lambda r^{n-1} G(u(r))dr \\ &\geq H(1) = u_r^2(1)/2 > 0. \end{aligned}$$

Thus we always have $H(r) > 0$ on $(0,1]$.

Since $Q(0) = 0$ and

$$Q'(u) = uf'(u) - f(u) = f(u)[K(u) - 1] \geq 0 \ \forall u \in [0,\beta),$$

we have $Q(u) \geq 0$ on $[0,\beta)$ and hence, by (5.2), $h'(r) > 0$ on (0,1), as required.

Denote

$$\mu(r) = 2/[K(u(r)) - 1].$$

Then $\mu(r)$ is strictly decreasing for $r \in (0,1]$, and by (5.1),

$$-\Delta v - \lambda f'(u)v = g(r), \quad g(r) = \lambda f(u)[K(u) - 1][\mu(r) - \mu].$$

With these preparations, we are now ready to show that ϕ does not change sign in B in case (i). We argue indirectly. Suppose $\phi(r)$ has a zero in (0,1). Then we can find $0 < t_1 \leq t_2 < 1$ such that

$$\phi(t_1) = 0, \phi(r) > 0 \ \forall r \in [0,t_1); \quad \phi(t_2) = 0, \phi(r) \neq 0 \ \forall r \in (t_2,1).$$

Now we choose $\mu = h(t_1)$ in $v = ru_r + \mu u$, and have two cases to consider:

$$\text{(a) } \mu(t_1) \geq \mu \text{ and (b) } \mu(t_1) < \mu.$$

We have

$$v(r) = ru_r + h(t_1)u = u[h(t_1) - h(r)] > v(t_1) = 0 \ \forall r \in [0,t_1), v(r) < 0 \ \ \forall r \in (t_1,1). \tag{5.4}$$

In case (a), we easily see $g(r) > 0$ on $(0,t_1)$, and hence, using $v(t_1) = 0$, we arrive at the following contradiction:

$$0 < \int_0^{t_1} g(r)\phi(r)r^{n-1}dr = \int_{B_{t_1}} [-\Delta v - \lambda f'(u)v]\phi = \int_{\partial B_{t_1}} v\phi_r = 0, \tag{5.5}$$

where we use $B_r = \{x \in R^n : |x| \leq r\}$.

In case (b), we may assume $\phi(r) > 0$ on $(t_2,1)$ for otherwise we can replace ϕ by $-\phi$. Moreover, one easily sees $g(r) < 0$ on $[t_1,1]$. Then by (5.4) and $\phi'(t_2) > 0 > \phi'(1)$, we also arrive at a contradiction:

$$0 > \int_{t_2}^1 g(r)\phi(r)r^{n-1}dr = \int_{B\setminus B_{t_2}} [-\Delta v - \lambda f'(u)v]\phi = \int_{\partial B_{t_2}} -v\phi_r + \int_{\partial B} v\phi_r > 0. \tag{5.6}$$

This proves the lemma for case (i).

Next we consider case (ii) where $u(0) > \beta$. Since $\beta = \infty$ when $m \geq 1$, we necessarily have $m < 1$ in this case. We can find $0 < r_1 < r_2 < 1$ uniquely determined by

$$u(r_1) = \beta,\ u(r_2) = \rho.$$

We first show $\phi(r) \neq 0$ on $(0, r_1]$. To this end, we choose $w(r) = u(r) - \rho$ as a test function. Clearly

$$-\Delta w - \lambda f'(u)w = \lambda q(u), \quad q(u) = f(u) - f'(u)(u - \rho).$$

We have

$$q'(u) = (\rho - u)f''(u),$$

which is positive on $(0, \rho)$, negative on (ρ, α), and positive on (α, ∞). Since $q(0) = \rho f'(0) = 0$, it follows that

$$q(u) \geq q(\alpha) = f(\alpha) - f'(\alpha)(\alpha - \rho) = 0\ \forall u > 0.$$

Hence

$$-\Delta w - \lambda f'(u)w = \lambda q(u) \geq 0 \text{ on } B.$$

If $\phi(r)$ has a zero in $(0, r_1]$, then we can find $t \in (0, r_1]$ such that $\phi(r) > 0$ on $[0, t)$ and $\phi(t) = 0$. Using $w(t) = u(t) - \rho > u(r_2) - \rho = 0$, we deduce

$$0 \leq \int_0^t \lambda q(u(r))\phi(r)r^{n-1}dr = \int_{B_t} [-\Delta w - \lambda f'(u)w]\phi = \int_{\partial B_t} w\phi_r < 0.$$

This contradiction finishes our proof for $\phi(r) \neq 0$ on $[0, r_1]$.

Next we suppose $\phi(r)$ changes sign in (0,1) and deduce a contradiction. Since $\phi(r) \neq 0$ in $[0, r_1]$, we can find $r_1 < t_1 \leq t_2 < 1$ such that

$$\phi(t_1) = 0, \phi(r) > 0\ \forall r \in [0, t_1); \quad \phi(t_2) = 0, \phi(r) \neq 0\ \forall r \in (t_2, 1).$$

As in case (i), we choose $\mu = h(t_1)$. Since $u(r) \leq \beta$ on $[r_1, 1]$, the above arguments for case (i) give

$$h'(r) > 0 \text{ on } (r_1, 1).$$

Hence

$$v(r) = u[h(t_1) - h(r)] > v(t_1) = 0\ \forall r \in [r_1, t_1);\ v(r) < 0\ \forall r \in (t_1, 1].$$

If $\mu(t_1) \geq \mu$, then since $\mu(r)$ is strictly decreasing on (0,1) and $K(u(r)) < 1$ on $[0, r_1)$, $K(u(r)) > 1$ on $(r_1, 1]$, we have

$$g(r) = \lambda f(u)[K(u) - 1][\mu(r) - \mu] > 0\ \forall r \in (r_1, t_1),$$

and by (5.1),

$$g(r) = \lambda f(u)[2 - \mu(K(u) - 1)] > 0\ \forall r \in [0, r_1].$$

Thus $g(r) > 0$ on $[0, t_1)$. Now we can deduce the same contradiction (5.5) as in case (i).

If $\mu(t_1) < \mu$, then

$$g(r) = \lambda f(u)[K(u) - 1][\mu(r) - \mu] < 0\ \forall r \in (t_1, 1],$$

and we arrive at the contradiction (5) as in case (i). This finishes the proof of Lemma 2.1. □

Acknowledgments. I wish to thank Professors Norm Dancer and Stuart Hastings for interesting discussions and valuable comments.

REFERENCES

[1] H. Amann, *Fixed point equations and nonlinear eigenvalue problems in ordered Banach spaces*, SIAM Rev., 18 (1976), pp. 620–709.
[2] R. Aris, *The Mathematical Theory of Diffusion and Reaction in Permeable Catalysts*, Oxford University Press, London, 1975.
[3] J. Bebernes and D. Eberly, *Mathematical Problems from Combustion Theory*, Springer-Verlag, New York, 1989.
[4] T. Boddington, P. Gray, and C. Robinson, *Thermal explosion and the disappearance of criticality at small activation energies: Exact results for the slab*, Proc. Roy. Soc. London Sect. A, 368 (1979), pp. 441–461.
[5] M.G. Crandall and P.H. Rabinowitz, *Bifurcation, perturbation of simple eigenvalues and linearized stability*, Arch. Rational Mech. Anal., 52 (1973), pp. 161–180.
[6] M.G. Crandall and P.H. Rabinowitz, *Some continuation and variational methods for positive solutions of nonlinear elliptic eigenvalue problems*, Arch. Rational Mech. Anal., 58 (1975), pp. 207–218.
[7] E.N. Dancer, *On the structure of solutions of an equation in catalysis theory when a parameter is large*, J. Differential Equations, 37 (1980), pp. 404–437.
[8] Y. Du, *Uniqueness, multiplicity and stability for positive solutions of a pair of reaction-diffusion equations*, Proc. Roy. Soc. Edinburgh Sect. A, 126 (1996), pp. 777–809.
[9] Y. Du and Y. Lou, *Proof of a conjecture for the perturbed Gelfand equation from combustion theory*, J. Differential Equations, to appear.
[10] Y. Du and Y. Lou, *Some uniqueness and exact multiplicity results for a predator-prey model*, Trans. Amer. Math. Soc., 349 (1997), pp. 2443–2475.
[11] Y. Du and Y. Lou, *S-shaped global bifurcation curve and Hopf bifurcation of positive solutions to a predator-prey model*, J. Differential Equations, 144 (1998), pp. 390–440.
[12] L. Erbe and M. Tang, *Uniqueness theorems of positive radial solutions of quasilinear elliptic equations in a ball*, J. Differential Equations, 138 (1997), pp. 351–379.
[13] J.E. Furter and J. Lopez-Gomez, *On the existence and uniqueness of coexistence states for the Lotka-Volterra competition model with diffusion and spatially dependent coefficients*, Nonlinear Anal., 25 (1995), pp. 363–398.
[14] J.M. Fraile, J. Lopez-Gomez, and J.C. Sabina de Lis, *On the global structure of the set of positive solutions of some semilinear elliptic boundary value problems*, J. Differential Equations, 123 (1995), pp. 180–212.
[15] B. Gidas, W.-M. Ni, and L. Nirenberg, *Symmetry and related properties via the maximum principle*, Comm. Math. Phys., 68 (1979), pp. 209–243.
[16] B. Gidas and J. Spruck, *A priori bounds for positive solutions of nonlinear elliptic equations*, Comm. Partial Differential Equations, 6 (1981), pp. 883–901.
[17] S.P. Hastings and J.B. McLeod, *The number of solutions to an equation from catalysis*, Proc. Roy. Soc. Edinburgh Sect. A, 101 (1985), pp. 15–30.
[18] P. Hess, *On uniqueness of positive solutions of nonlinear elliptic boundary value problems*, Math. Z., 154 (1977), pp. 17–18.
[19] D.D. Joseph and T.S. Lundgren, *Quasilinear Dirichlet problems driven by positive sources*, Arch. Rational Mech. Anal., 49 (1973), pp. 241–269.
[20] A.K. Kapila and B.J. Matkowsky, *Reactive-diffusive systems with Arrhenius kinetics: Multiple solutions, ignition and extinction*, SIAM J. Appl. Math., 36 (1979), pp. 373–389.
[21] A.K. Kapila, B.J. Matkowsky, and J. Vega, *Reactive-diffusive systems with Arrhenius kinetics: Peculiarities of the spherical geometry*, SIAM J. Appl. Math., 38 (1980), pp. 382–401.
[22] P. Korman and Y. Li, *On the exactness of an S-shaped bifurcation curve*, Proc. Amer. Math. Soc., 127 (1999), pp. 1011–1020.
[23] P. Korman, Y. Li, and T. Ouyang, *An exact multiplicity result for a class of semilinear equations*, Comm. Partial Differential Equations, 22 (1997), pp. 661–684.
[24] C.S. Lin and W.-M. Ni, *A counterexample to the nodal domain conjecture and a related semilinear equation*, Proc. Amer. Math. Soc., 102 (1988), pp. 271–277.
[25] J. Lopez-Gomez, *Varying bifurcation diagrams of positive solutions for a class of indefinite superlinear boundary value problems*, Trans. Amer. Math. Soc., 352 (2000), pp. 1825–1858.

[26] K. Nagasaki and T. Suzuki, *Spectral and related properties about the Emden-Fowler equation $-\Delta u = \lambda e^u$ on circular domains*, Math. Ann., 299 (1994), pp. 1–15.
[27] W.-M. Ni and R.D. Nussbaum, *Uniqueness and nonuniqueness for positive radial solutions of $\Delta u + f(u, r) = 0$*, Comm. Pure Appl. Math., 38 (1985), pp. 67–108.
[28] T. Ouyang and J. Shi, *Exact multiplicity of positive solutions for a class of semilinear problems*, J. Differential Equations, 146 (1998), pp. 121–156.
[29] S.V. Parter, *Solutions of a differential equation arising in chemical reactor processes*, SIAM J. Appl. Math., 26 (1974), pp. 687–716.
[30] P.H. Rabinowitz, *Minimax Methods in Critical Point Theory with Applications to Differential Equations*, CBMS Reg. Conf. Ser. Math. 65, AMS, Providence, RI, 1986.
[31] G.C. Wake, T. Boddington, and P. Gray, *Thermal explosion and the disappearance of criticality in systems with distribution tempratures.* IV*. Rigorous bounds and their practical relevance*, Proc. Roy. Soc. London Sect. A, 425 (1989), pp. 285–289.
[32] S.-H. Wang, *On S-shaped bifurcation curves*, Nonlinear Anal., 22 (1994), pp. 1475–1485.
[33] S.-H. Wang, *Rigorous analysis and estimates of S-shaped bifurcation curves in a combustion problem with general Arrhenius reaction-rate laws*, Proc. Roy. Soc. London Sect. A, 454 (1998), pp. 1031–1048.
[34] S.-H. Wang and F.-P. Lee, *Bifurcation of an equation from catalysis theory*, Nonlinear Anal., 23 (1994), pp. 1167–1187.
[35] J. Wei, *Exact multiplicity for some nonlinear elliptic equations in balls*, Proc. Amer. Math. Soc., 125 (1997), pp. 3235–3242.

SIAM J. MATH. ANAL.
Vol. 32, No. 4, pp. 734–759

CONSTANT PRINCIPAL STRAIN MAPPINGS ON 2-MANIFOLDS*

MARTIN CHUAQUI† AND JULIAN GEVIRTZ†

Abstract. We study mappings between Riemannian 2-manifolds which have constant principal stretching factors (cps-mappings). Such mappings f can be described in terms of the relationship between the geodesic curvature of the curves of principal strain at p and that of their images at $f(p)$. In the context of local coordinates this relationship takes the form of a nonlinear hyperbolic system, the blow-up properties of which depend on the Gaussian curvatures of the two manifolds. We use the theory of such systems to study global existence when both manifolds are the hyperbolic plane $\mathbb{H}^2$ and obtain a simple description of all cps-mappings of $\mathbb{H}^2$ onto itself. We also obtain a distortion result for disks in $\mathbb{H}^2$ as well as some nonexistence results for cps-mappings of the Euclidean plane onto certain classes of manifolds. In addition, our treatment of cps-mappings in $\mathbb{H}^2$ yields, virtually as a corollary, a generalization of a theorem of Epstein to the effect that a curve in hyperbolic n-space whose geodesic curvature is bounded by 1 must be simple.

Key words. constant principal strains, hyperbolic system, hyperbolic plane

AMS subject classifications. Primary, 35L45, 35L60, 53B20, 53C99; Secondary, 73G99

PII. S0036141099352534

1. Introduction.

Consider a thin liquid film which upon solidification acquires a cryptocrystalline structure; that is, at each point a suitably oriented infinitesimal square of the original liquid becomes an (again, suitably oriented infinitesimal) rectangular crystal whose side lengths are constant multiples of the side length of the square. Such a process produces a deformation of the surface originally formed by the liquid, and in this paper we examine the class of deformations—those having constant principal strains—that can be realized in this manner. It turns out that the associated mappings are governed by hyperbolic systems of partial differential equations, a circumstance which in retrospect is not surprising since one would expect that singularities, in higher derivatives of the deformation, for example, propagate along the sides of the microscopic crystals, that is, along the associated curves of principal strain. This hyperbolicity in conjunction with the additional element of nonlinearity underlies most of what follows.

To give an idea of some of the relevant issues, we briefly describe the situation in the planar context (see [Ge1] for further details). Let $0 < m_1 < m_2$. A differentiable, orientation preserving mapping f of a domain $U \subset \mathbb{R}^2$ into $\mathbb{R}^2$ has constant principal stretches m_1, m_2 if there are functions $\theta, \overline{\theta}$ on U such that its Jacobian J_f satisfies

$$J_f = T(-\overline{\theta})S(m_1, m_2)T(\theta), \tag{1.1}$$

where

$$T(\theta) = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \quad \text{and} \quad S(m_1, m_2) = \begin{bmatrix} m_1 & 0 \\ 0 & m_2 \end{bmatrix}.$$

Throughout, such f will be called (m_1, m_2)-mappings, or less specifically cps-mappings ("cps" for constant principal strain). This direct manner of expressing the

*Received by the editors February 17, 1999; accepted for publication June 8, 2000; published electronically November 10, 2000. This work was partially supported by Fondecyt grant 1971055.
http://www.siam.org/journals/sima/32-4/35253.html
†Facultad de Matemáticas, P. Universidad Católica de Chile, Casilla 306, Santiago 22, Chile (mchuaqui@mat.puc.cl, jgevirtz@mat.puc.cl).

condition that a mapping has constant principal stretches m_1, m_2 turns out to be rather uninformative, it being far better to work with the compatibility conditions for a matrix function to be a Jacobian; for this reason one adds the additional hypothesis that J_f be locally Lipschitz continuous on U. (See the first paragraph of section 5 for comments about this regularity assumption.) A straightforward calculation shows that a necessary and sufficient condition that locally Lipschitz functions θ and $\overline{\theta}$ give the Jacobian of an (m_1, m_2)-mapping (in a simply connected domain) via the formula (1.1) is that

$$D_1(m_1\theta - m_2\overline{\theta}) = 0 \quad \text{and} \quad D_2(m_2\theta - m_1\overline{\theta}) = 0 \tag{1.2}$$

hold almost everywhere (a.e.), where D_1 and D_2 denote differentiation in the directions $e^{i\theta}$ and $ie^{i\theta}$, respectively. These equations relate the curvatures of the curves (to be referred to henceforth as i-characteristics) along which the stretching factor is m_i and their images. Indeed, if the curvature of the former at $p \in U$ is κ_i and that of the latter at $f(p)$ is $\overline{\kappa}_i$, then (1.2) simply says that $\overline{\kappa}_i = \kappa_i/m_j$, where $\{i, j\} = \{1, 2\}$. These equations constitute a genuinely nonlinear diagonal hyperbolic system for the pair of functions θ, $\overline{\theta}$, so that, in light of a general principle established by Lax [L], one expects cps-mappings to display a marked tendency to form singularities. Specifically, the blow-up law for system (1.2) says, in the case of sufficiently differentiable mappings (and actually for all cps-mappings in the appropriate weak sense), that at each point p the derivative of κ_i in the direction of the j-characteristic through p and toward the concave side of the i-characteristic through this point is κ_i^2, from which it follows at once that the curvatures of both of the characteristics of f at p are bounded above by $1/\mathrm{dist}(p, \partial U)$. Two immediate consequences of this are (i) a cps-analogue of Liouville's theorem—the only cps-mappings of the entire plane onto itself are affine and (ii) the compactness of the class of all (m_1, m_2)-mappings of U into $\mathbb{R}^2$ with respect to the topology of uniform convergence of the first-order derivatives on compact subsets. This blow-up principle also allows one to show that the radius of the largest concentric subdisk of the unit disk Δ whose image under all (m_1, m_2)-mappings $f : \Delta \to \mathbb{R}^2$ is convex is $\left(\frac{m_1}{m_2}\right)^2$. In fact, in conjunction with (1.2) the growth law for the κ_i plays a decisive role in the analysis of other aspects of cps-mappings and of the intimately related "principal strain line inclination function" θ (whose integral curves together with their orthogonal trajectories form what is known in plasticity and optimum structure theory—see [Hil] and [He]—as Hencky–Prandtl nets), such as boundary behavior [Ge3], [Ge4], the nature and distribution of isolated singularities [Ge3], and the determination of all cps-self-homeomorphisms of certain domains [Ge4]. A number of these properties of cps-mappings are strikingly similar to their conformal analogues.

In the present paper we examine some of these issues in the context of 2-dimensional manifolds. We begin in section 2 by establishing the counterparts of (1.2) and the blow-up law, whose formal derivations are somewhat more involved than in the planar case. In section 3 we discuss the analytic details necessary to deal with questions of global existence and behavior, and in addition analyze the relationship between cps-mappings and a generalization of Hencky–Prandtl nets in the constant Gaussian curvature context; more than anything these considerations involve appropriate rewriting of the equations derived in section 2 in coordinate form so as to make manifest the exact nature of the underlying hyperbolicity. In section 4 we apply the results of section 3 first to show that in certain situations there exist no globally defined cps-mappings and then, in the special case of the hyperbolic plane $\mathbb{H}^2$, to do the

following: (i) completely describe the (wide) class of cps-mappings of $\mathbb{H}^2$ onto itself, (ii) prove a generalization of a theorem of Epstein [E1], [E2] about the curvature of self-intersecting curves in hyperbolic n-space $\mathbb{H}^n$, and (iii) derive an analogue for $\mathbb{H}^2$ of the planar radius of convexity result mentioned in the preceding paragraph.

In the planar context one could consider in addition to cps-mappings other similarly defined classes such as the one consisting of mappings with Jacobian of the form

$$J_f = T(-\overline{\theta})S(m_1(\theta,\overline{\theta}), m_2(\theta,\overline{\theta}))T(\theta)$$

for any given pair of everywhere distinct positive functions $m_1(\theta,\overline{\theta}), m_2(\theta,\overline{\theta})$ of period π in each variable (that is, mappings for which the principal strains are given functions of the directions of the principal strain lines and their images). Such a generalization is not possible in context of Riemannian 2-manifolds owing to the absence of an absolute reference direction. Indeed, since the principal stretches (and combinations of them) are the only intrinsically definable first-order parameters associated with a mapping between manifolds, in this context there are only two natural classes of mappings defined by point-independent conditions on their Jacobians: conformal mappings and (m_1, m_2)-mappings. (We are considering here only families of mappings for which, loosely speaking, the set of possible Jacobians at each point is governed by two parameters.) For this reason, cps-mappings constitute a natural object of study above and beyond their interpretation as deformations arising in certain physical situations.

2. Formal considerations. Let V and $\overline{V}$ be C^∞ Riemannian 2-manifolds, both metric tensors being denoted by $\langle\cdot,\cdot\rangle$, which we sometimes subscript with V or $\overline{V}$ for additional clarity. Let $U \subset V$ be a domain. The principal stretches (henceforth to be called principal strains in slight abuse of accepted terminology) of a mapping $f : U \to \overline{V}$ at a point $p \in V$ at which the Jacobian transformation $J_f(p)$ is nonsingular are the square roots of the eigenvalues of the transformation $J_f^*(p)J_f(p)$ of the tangent space of V at p onto itself. Let $U \subset V$ be a domain and m_1, m_2 be distinct positive constants. Then $f : U \to \overline{V}$ is an (m_1, m_2)-mapping if J_f is locally Lipschitz continuous and the principal strains of f are everywhere given by the pair (m_1, m_2). As one can imagine from what was said above about the planar case, the direct expression of this condition as a nonlinear 2×2 system of partial differential equations in terms of local coordinate systems for V and $\overline{V}$ is not very revealing, although as we shall explain in section 3 a small amount of information can be gleaned from it. Here also it is much more appropriate to consider a derived higher order system, specifically a second-order one—which has an elegant coordinate-free formulation—in which the geometric structures of V and $\overline{V}$ present themselves in a most transparent way.

In dealing with the differential geometric aspects we shall, apart from minor variations, adhere to the notation of Hicks [Hic] . In general, the counterpart for $\overline{V}$ of any object A associated with V will be denoted by $\overline{A}$. The Lie bracket of two vector fields X_1, X_2 will be denoted as usual by $[X_1, X_2]$. It is clear that if $U \subset V$ is a simply connected domain, then $f : U \to \overline{V}$ is an (m_1, m_2)-mapping if and only if its Jacobian J_f is locally Lipschitz continuous and there exist locally Lipschitz continuous fields X_1, X_2 on U such that $\langle X_i, X_j\rangle = \delta_{ij}$ and $\langle J_f X_i, J_f X_j\rangle = m_i m_j \delta_{ij}$. The fields X_1, X_2 are *principal direction fields* for f.

The unit vector $J_f X_i/m_i$ will be denoted by $\overline{X}_i$. The covariant derivative in the direction X of the vector field Y will be denoted by $D_X Y$. In addition, D_{X_i} ($D_{\overline{X}_i}$) will be abbreviated by D_i ($\overline{D_i}$), and the same symbols $D_X\alpha$, $D_i\alpha$ will be used to denote the derivative of the scalar function α in the corresponding directions. We shall use

the following facts (see [Hic]). If $f : U \to \overline{V}$ is a diffeomorphism and X, Y, and Z are vector fields on V, then

$$J_f[X, Y] = [J_f X, J_f Y], \tag{2.1}$$

$$D_X Y - D_Y X = [X, Y], \tag{2.2}$$

and

$$D_X\langle Y, Z\rangle = \langle D_X Y, Z\rangle + \langle Y, D_X Z\rangle. \tag{2.3}$$

Furthermore, if Y is a vector field and α, β are scalar functions, then

$$D_X(\alpha Y) = (D_X\alpha)Y + \alpha D_X Y \tag{2.4}$$

and

$$D_{\alpha X+\beta Z}(Y) = \alpha D_X Y + \beta D_Z Y. \tag{2.5}$$

Let $\{X_1, X_2\}$ be an orthonormal pair of locally Lipschitz vector fields on some domain U in V. The covariant derivative $D_l X_k$ exists a.e., and the equations appearing in this paragraph hold a.e. in U. As a consequence of (2.3) we have that

$$0 = D_l\langle X_j, X_k\rangle = \langle D_l X_j, X_k\rangle + \langle X_j, D_l X_k\rangle$$

so that

$$\langle D_l X_j, X_j\rangle = 0 \quad \text{and} \quad \langle D_l X_j, X_k\rangle = -\langle D_l X_k, X_j\rangle \tag{2.6}$$

and with the convention that $\{i, j\} = \{1, 2\}$, which will be in force throughout, this means that there are locally bounded measurable scalar functions κ_i such that

$$D_i X_i = \kappa_i X_j \ \text{ and } \ D_i X_j = -\kappa_i X_i. \tag{2.7}$$

At a point p at which it exists (and it does so a.e. on U), $\kappa_i(p)$ is the geodesic curvature of the integral curve through p of the field X_i. Now consider the pairs of orthonormal fields $\{X_1, X_2\}$ and $\{\overline{X}_1, \overline{X}_2\}$ associated with an (m_1, m_2)-mapping $f : U \to \overline{V}$. It follows from (2.2) and (2.7) that

$$[X_i, X_j] = D_i X_j - D_j X_i = \kappa_j X_j - \kappa_i X_i \tag{2.8}$$

so that

$$\kappa_j = \langle [X_i, X_j], X_j\rangle.$$

By (2.8) and (2.1), which may be applied since f is a local diffeomorphism,

$$\begin{aligned}
\overline{\kappa}_j &= \langle [\overline{X}_i, \overline{X}_j], \overline{X}_j\rangle = \langle [J_f X_i/m_i, J_f X_j/m_j], \overline{X_j}\,\rangle = \langle [J_f X_i, J_f X_j], \overline{X_j}\,\rangle/m_i m_j \\
&= \langle J_f[X_i, X_j], \overline{X_j}\,\rangle/m_i m_j = \langle J_f(\kappa_j X_j - \kappa_i X_i), \overline{X_j}\,\rangle/m_i m_j \\
&= \langle \kappa_j J_f X_j - \kappa_i J_f X_i, \overline{X_j}\,\rangle/m_i m_j \\
&= \langle \kappa_j m_j \overline{X}_j - \kappa_i m_i \overline{X}_i, \overline{X_j}\,\rangle/m_i m_j = \kappa_j/m_i.
\end{aligned}$$

We thus have the fundamental *curvature equations*

$$\overline{\kappa}_j = \kappa_j/m_i \text{ a.e. in } U, \qquad j = 1, 2. \tag{2.9}$$

We next consider how the curvatures change as we move along characteristics, and for the time being we shall assume that the mapping in question is of class C^3. (We shall explain in section 3—see Theorem 3.2—in what way this additional regularity requirement is in fact superfluous.) We use the fact that the Gaussian curvature of a 2-dimensional manifold V at a point p is given by $\langle R(X,Y)Y, X\rangle$ for all orthonormal pairs X, Y of vectors in the tangent space of V at p, where

$$R(X,Y)Y = D_X D_Y Y - D_Y D_X Y - D_{[X,Y]}Y.$$

In particular we have from (2.7) and (2.8)

$$\begin{aligned} R(X_1, X_2)X_2 &= D_1 D_2 X_2 - D_2 D_1 X_2 - D_{[X_1,X_2]}X_2 \\ &= D_1(\kappa_2 X_1) + D_2(\kappa_1 X_1) - D_{\kappa_2 X_2 - \kappa_1 X_1}X_2, \end{aligned}$$

so that upon taking into account (2.4), (2.5), and (2.7) again, we have

$$R(X_1, X_2)X_2 = \kappa_1\kappa_2 X_2 + (D_1\kappa_2)X_1 - \kappa_1\kappa_2 X_2 + (D_2\kappa_1)X_1 - \kappa_2^2 X_1 - \kappa_1^2 X_1.$$

Thus, if K and $\overline{K}$ denote Gaussian curvature on V and $\overline{V}$, we have

$$K = \langle R(X_1, X_2)X_2, X_1\rangle = D_1\kappa_2 + D_2\kappa_1 - \kappa_2^2 - \kappa_1^2 \tag{2.10}$$

and

$$\overline{K} = \langle R(\overline{X}_1, \overline{X}_2)\overline{X}_2, \overline{X}_1\rangle = \overline{D}_1\overline{\kappa}_2 + \overline{D}_2\overline{\kappa}_1 - \overline{\kappa}_2^2 - \overline{\kappa}_1^2. \tag{2.11}$$

In light of the fundamental relations (2.9) and the fact that $\overline{X}_i = J_f X_i/m_i$, it then follows that $\overline{D}_i\overline{\kappa}_j(f(p)) = (D_i\kappa_j(p))/m_i^2$, so that (2.11) may be written as

$$\overline{K} = (D_1\kappa_2)/m_1^2 + (D_2\kappa_1)/m_2^2 - \kappa_2^2/m_1^2 - \kappa_1^2/m_2^2. \tag{2.12}$$

Upon solving the linear system for $D_1\kappa_2$ and $D_2\kappa_1$ given by (2.10) and (2.12), we obtain

$$D_j\kappa_i = \kappa_i^2 + c_i, \qquad i = 1, 2, \tag{2.13}$$

where

$$c_i = m_j^2 \frac{m_i^2\overline{K} - K}{m_i^2 - m_j^2}. \tag{2.14}$$

We emphasize that when these *blow-up equations* (2.13) are written out fully in coordinate form the functions giving the mapping itself appear as arguments of $\overline{K}$, so that they do not in general characterize the net of principal strain lines in an intrinsic fashion. Although they purport to tell us something about how far along a characteristic from a given point a singularity—a point where the mapping fails to be locally Lipschitz—must lie, their content in this regard is meaningless unless one has information about K and $\overline{K}$. For this reason, the most interesting cases by far are those in which at least one of these curvatures is constant.

Given an orthonormal pair of fields X_1, X_2 on $U \subset V$ we refer to arcs of the integral curves of the field X_k as k-arcs. A domain $Q \subset U$ will be said to be a *characteristic quadrilateral* of X_1, X_2 (or of an associated cps-mapping) if ∂Q is a Jordan curve lying in D containing four points a, b, c, d occurring in that order when ∂D is traversed (in one direction or the other) and such that ab and cd are i-arcs, and bc and da are j-arcs. For such a Q we denote by Q_i^+ the i-side (i.e., ab or cd) along which X_j points toward the inside of Q. This i-side of Q will be referred to as the positive i-side. The other, negative, i-side will be denoted by Q_i^-. For an i-arc C we write

$$\Delta(C) = \int_C \kappa_i ds,$$

the unoriented arc length integral of κ_i along C. Let $U \subset V$ be simply connected, and let $f : U \to \overline{V}$ be an (m_1, m_2)-homeomorphism. For each characteristic quadrilateral $Q \subset U$ the positive sides of Q are mapped onto the positive sides of the image quadrilateral $\overline{Q}$. Because the exterior angles of a characteristic quadrilateral are all $\pi/2$, the Gauss–Bonnet formula says

$$\Delta(Q_1^+) - \Delta(Q_1^-) + \Delta(Q_2^+) - \Delta(Q_2^-) = -\int_Q K dA. \tag{2.15}$$

However, the cps-conditions and (2.9) together imply that

$$\Delta(\overline{Q}_i^\sigma) = \frac{m_i}{m_j}\Delta(Q_i^\sigma), \qquad i = 1, 2,\ \sigma = +, -,$$

so that application of the Gauss–Bonnet formula to $\overline{Q}$ gives

$$\frac{m_1}{m_2}\big(\Delta(Q_1^+) - \Delta(Q_1^-)\big) + \frac{m_2}{m_1}\big(\Delta(Q_2^+) - \Delta(Q_2^-)\big) = -m_1 m_2 \int_Q \overline{K}(f) dA. \tag{2.16}$$

Upon solving the system (2.15), (2.16) for the $\Delta(Q_i^+) - \Delta(Q_i^-)$, we obtain

$$\Delta(Q_i^+) - \Delta(Q_i^-) = -\int_Q c_i dA \tag{2.17}$$

for every closed characteristic quadrilateral $Q \subset U$, where c_i is given in (2.14). Although we have only shown that (2.17) holds for quadrilaterals on whose closure f is one-to-one, these equations can easily be seen to hold for any characteristic quadrilateral by the standard process of breaking them up into smaller quadrilaterals. We note that in light of (2.15) and the fact that $c_1 + c_2 = K$ the validity of (2.17) with one value of i implies it with the other one.

In the planar context a Hencky–Prandtl (HP) net on a simply connected domain D consists of two mutually orthogonal one-parameter families of curves covering D with the property that for any two fixed curves C_1, C_2 belonging to one of the families, the change in the inclination of the tangent is the same along all subarcs of curves of the other family which join a point of C_1 to a point of C_2. For simply connected domains, an orthogonal pair of curve families is an HP net if and only if it is the net of principal strain lines of a cps-mapping. This gives an intrinsic characterization of principal strain lines that, unlike one based (2.13), does not make reference to third-order derivatives. In order to obtain such an intrinsic characterization in the

nonplanar context, one needs to assume that the curvature $\overline{K}$ of the image manifold $\overline{V}$ is constant, and in order to avoid a clumsy formulation as well as to preserve the symmetry of the discussion, we shall assume that the curvature K of V is constant as well. Thus for such a V we will say that two mutually orthogonal locally Lipschitz unit vector fields X_1, X_2 on a simply connected domain $U \subset V$ are an $(m_1, m_2, \overline{K})$-HP pair if either of the equations

$$\Delta(Q_i^+) - \Delta(Q_i^-) = -c_i A(Q),$$

where $A(Q)$ is the area of Q, is satisfied for all relevant quadrilaterals; here, of course, the curvatures κ_i are defined by the first equation in (2.7). Thus we have derived the following.

THEOREM 2.1. *If V and $\overline{V}$ have constant Gaussian curvature K and $\overline{K}$, respectively, and $f : V \to \overline{V}$ is an (m_1, m_2)-mapping, then the corresponding principal fields X_1, X_2 are an $(m_1, m_2, \overline{K})$-HP pair.*

In the next section (see Theorem 3.4) we show that, conversely, given an $(m_1, m_2, \overline{K})$-HP pair on a simply connected domain U in V, there is an (m_1, m_2)-mapping f of U into a manifold $\overline{V}$ with constant Gaussian curvature $\overline{K}$, and that this mapping is unique up to rigid motions in $\overline{V}$.

3. Analytic considerations. In investigating cps-mappings two fundamental directions are to be pursued. On the one hand, one would like to say something about the global behavior of all possible cps-mappings of a given domain, that is, to develop some elements of a distortion theory for such mappings. This aspect of the theory is to be based on the three fundamental relations derived in the preceding section: the curvature equations, the blow-up equations, and the HP property, and an example will be discussed in section 4. On the other hand, one should also be able to manufacture such mappings, that is, to construct solutions to the corresponding differential equations, and this is the point we address in this section.

The most straightforward approach is that of DeTurck and Yang [DY] in which one considers the differential equations which state that the eigenvalues of the transformation $J_f^*(p)J_f(p)$ (of the tangent space at p onto itself) are the m_i^2. Specifically, we consider coordinates (u_1, u_2) and $(\overline{u}_1, \overline{u}_2)$ for neighborhoods $U, \overline{U}$ in $V, \overline{V}$, respectively. For convenience we further assume that $U = \{(u_1, u_2) \,|\, |u_1|, |u_2| < \epsilon\}$. In terms of these coordinate systems let $(f_1, f_2) = f : U \to \overline{U}$ be an (m_1, m_2)-mapping for which the length change produced by f on the arc corresponding to $u_2 = 0$ is everywhere strictly between m_1 and m_2. DeTurck and Yang showed that there are four pairs of real-analytic functions F_k^σ, $1 \le \sigma \le 4$, $k = 1, 2$, of twelve variables such that for one of the four values of σ,

$$\frac{\partial f_k}{\partial u_2}(u) = F_k^\sigma \left(\frac{\partial f}{\partial u_1}, m_1, m_2, G(u), \overline{G}(f(u)) \right), \qquad k = 1, 2,$$

where each of G and $\overline{G}$ stands for the four elements of the metric tensors of V and $\overline{V}$ evaluated as indicated. Each of these systems makes the required statement about the eigenvalues of $J_f^*(p)J_f(p)$, and that there are four of them is simply a reflection of the fact that for any given m strictly between m_1 and m_2, and any nonzero $e \in \mathbb{R}^2$, there are four distinct linear transformations $T : \mathbb{R}^2 \to \mathbb{R}^2$ with principal stretches m_1, m_2 for which $Te = me$ (two orientation preserving and two orientation reversing). Conversely, in the analytic category the Cauchy–Kowalewski theorem implies that for each of these four systems the initial value problem $f(u_1, 0) = f_0(u_1)$ has a unique local

solution provided that along the curve $u_2 = 0$ the given initial mapping f_0 changes arc length by factors lying strictly between m_1 and m_2. DeTurck and Yang made the additional very important observation that the linearizations of these four systems are diagonal hyperbolic, and this allowed them to deduce local existence in the C^∞ category. (Their work is actually considerably more general in that it deals with mappings with distinct principal strains on manifolds of arbitrary dimension.) In [Ge2] we dubbed the four Cauchy problems collectively the *DeTurck–Yang initial value problem*, a term we shall employ in what follows to refer to any one of them.

This approach to the construction of cps-mappings as solutions to first-order systems, however, throws no light on global existence because it reveals nothing about how, where, or why singularities form. Information of this nature is, on the other hand, implicit in the blow-up equations and can be put to use by basing the construction of cps-mappings either directly on them or, better still, on the analytically simpler system of curvature equations. We pursue this latter option, but because there are only two distinct characteristics and we are interested in working with the absolutely minimal condition of locally Lipschitz continuity of J_f, we do so via the method of characteristic coordinates. We begin by deriving the necessary equations.

Let U be a (small) neighborhood in V and let (u_1, u_2) be local coordinates for U; in what follows we freely identify points $p \in U$ with the corresponding $(u_1, u_2) \in \mathbb{R}^2$. We denote by $e_k = e_k(p)$ the Euclidean unit vectors at $p \in U$. A right-hand orthonormal pair (with respect to the metric of V) of vectors X_1, X_2 at $u \in U$ is completely specified by the inclination θ of X_1 to the positive u_1-axis. In other words, there are functions $\alpha_k^{(i)}(u, \theta)$ such that in terms of θ

$$X_i = \sum_{k=1}^{2} \alpha_k^{(i)}(u, \theta) e_k = F_i(u, \theta). \tag{3.1}$$

If we are dealing with a real-analytic manifold, then the $\alpha_k^{(i)}$ are, of course, real-analytic. In the discussion to follow, β will denote specific but not explicitly calculated (vector- or scalar-valued) functions of arguments to be indicated; these functions will easily be seen to be real-analytic when we are in that category and to be independent of the particular fields X_1, X_2. It is to be borne in mind that the functions denoted by this symbol may change from line to line and that the symbol D_i ($\overline{D}_i$) is used to denote both differentiation of scalar functions and covariant differentiation of vector fields in the direction X_i ($\overline{X}_i$). In the calculations to follow we use covariant differentiation rules (2.4) and (2.5). We have

$$\begin{aligned} D_i X_i = D_i \left(\sum_{k=1}^{2} \alpha_k^{(i)}(u, \theta) e_k \right) &= \sum_{k=1}^{2} (D_i \alpha_k^{(i)}(u, \theta)) e_k + \beta(u, \theta) \\ &= (D_i \theta\,) \sum_{k=1}^{2} \frac{\partial \alpha_k^{(i)}(u, \theta)}{\partial \theta} e_k + \beta(u, \theta). \end{aligned}$$

Since $\kappa_i = \langle D_i X_i, X_j \rangle$, it follows that

$$\kappa_i = \left\langle \sum_{k=1}^{2} \frac{\partial \alpha_k^{(i)}}{\partial \theta} e_k, X_j \right\rangle D_i \theta + \beta(u, \theta) = P_i(u, \theta) D_i \theta + \beta(u, \theta), \tag{3.2}$$

where $P_i(u, \theta) = \langle \frac{\partial X_i}{\partial \theta}, X_j \rangle$. Since $\frac{\partial \langle X_i, X_j \rangle}{\partial \theta} = 0$, it follows that

$$P_j(u, \theta) = -P_i(u, \theta). \tag{3.3}$$

Because X_1 is of the form $\beta(u,\theta)(\cos\theta e_1 + \sin\theta e_2)$,

$$P_1 = \left\langle \frac{\partial X_1}{\partial \theta}, X_2 \right\rangle = \left\langle \frac{\partial \beta}{\partial \theta}(u,\theta)(\cos\theta e_1 + \sin\theta e_2) + \beta(u,\theta)(-\sin\theta e_1 + \cos\theta e_2), X_2 \right\rangle$$
$$= \beta(u,\theta)\langle -\sin\theta e_1 + \cos\theta e_2, X_2\rangle \neq 0,$$

since $-\sin\theta e_1 + \cos\theta e_2$ is not a multiple of X_1. Thus, in light of (3.3) we have

$$D_i\theta = R_i(u,\theta)\kappa_i + S_i(u,\theta), \tag{3.4}$$

where $R_i(p,\theta)$ and $S_i(p,\theta)$ are functions which for given V depend only on the arguments $p \in V$ and θ.

Let X_1, X_2 be an orthonormal pair of Lipschitz continuous fields on $U \subset V$ and let $S_\epsilon = \{(t_1,t_2) : -\epsilon < t_1, t_2 < \epsilon\}$. A bi-Lipschitz homeomorphism $u : S_\epsilon \to U$ is a characteristic coordinate mapping if each segment $t_i =$ constant is carried onto a j-characteristic. The Lipschitz continuity of the X_i imply that such mappings exist locally. With reference to such a mapping, in what follows Y_i will denote the tangent field $J_u e_i$, where the e_i are the Euclidean unit vector fields on S_ϵ; more concretely, $(D_{Y_i} w)(u(t_1,t_2)) = \partial w(u(t_1,t_2))/\partial t_i$ for scalar functions w. Obviously, $[Y_i, Y_j] = 0$. Furthermore, we define $y_i(t_1,t_2)$ by

$$Y_i(u(t_1,t_2)) = y_i(t_1,t_2)X_i(u(t_1,t_2)) = y_i(t_1,t_2)F_i(u(t_1,t_2),\theta(u(t_1,t_2))), \tag{3.5}$$

where F_i is the vector-valued function appearing in (3.1). Note that Y_i and y_i only exist a.e. on $u(S_\epsilon)$ and S_ϵ, respectively.

Assuming for the moment that u has enough regularity for the calculations to make sense, we have from the rules (2.4) and (2.5) of covariant differentiation together with (2.7) that

$$D_{Y_j} Y_i = (D_{Y_j} y_i)X_i - \kappa_j y_i y_j X_j \tag{3.6a}$$

and by symmetry that

$$D_{Y_i} Y_j = (D_{Y_i} y_j)X_j - \kappa_i y_j y_i X_i. \tag{3.6b}$$

(In these formulas $y_k = y_k(u^{-1}(p))$.)

Rule (2.2) and the fact $[Y_i, Y_j] = 0$ imply equality of the right-hand sides of (3.6a) and (3.6b) from which it follows that

$$\frac{\partial y_i}{\partial t_j} = -\kappa_i\, y_i y_j. \tag{3.7}$$

For pairs of functions $\eta = (\eta_1, \eta_2)$, $y = (y_1, y_2)$ we define

$$I_1(\eta, y) = \int_0^{t_2} \eta_1(t_1,t)y_1(t_1,t)y_2(t_1,t)dt, \tag{3.8a}$$

$$I_2(\eta, y) = \int_0^{t_1} \eta_2(t,t_2)y_1(t,t_2)y_2(t,t_2)dt. \tag{3.8b}$$

We need the following lemma which says in what sense (3.7) holds in general.

LEMMA 3.1. *For almost all* $t_1 \in (-\epsilon, \epsilon)$, y_1 *as a function of* t_2 *satisfies*

$$y_1(t_1, t_2) = y_1(t_1, 0) - I_1(\kappa, y) \tag{3.9}$$

for almost all $t_2 \in (-\epsilon, \epsilon)$ *and analogously for* y_2. *(Here,*$\kappa_i = \kappa_i(u(t_1, t_2))$.*)*

Proof. It is enough to show that this is the case for sufficiently small ϵ, since one can then patch together small squares to conclude that it is so in the original square. If u is a characteristic coordinate mapping, then so is $v(t_1, t_2) = u(f_1(t_1), f_2(t_2))$ for any pair of bi-Lipschitz functions f_1, f_2. If w_k is the counterpart of y_k for v, then

$$w_k(t_1, t_2) = y_k(f_1(t_1), f_2(t_2)) f_k'(t_k),$$

from which one sees that it is sufficient to prove the statement in the case that y_1 and y_2 are identically 1 on the lines $t_2 = 0$ and $t_1 = 0$, respectively. By working with a sequence of smooth approximations to the θ which gives X_1, X_2, we can approximate the pair X_1, X_2 by sequences $X_1^{(n)}, X_2^{(n)}$of orthonormal C^∞ fields which converge uniformly to the X_i in a neighborhood U of the closure of $u(S_\epsilon)$, for which the corresponding curvatures $\kappa_i^{(n)}$ are uniformly bounded and converge to the κ_i in $L^1(S_\epsilon)$, and such that $X_i^{(n)}(u(0,0)) = X_i(u(0,0))$. We consider the corresponding characteristic coordinate mappings $u^{(n)}$ with corresponding $Y_i^{(n)}$ and $y_i^{(n)}$, where $y_i^{(n)}$ is identically 1 on the line $t_j = 0$. Since the $y_i^{(n)}$ are smooth they satisfy (3.7) and consequently

$$y_i^{(n)} = 1 - I_i(\kappa_i^{(n)}, y^{(n)}), \qquad i = 1, 2. \tag{3.10}$$

Clearly, $u^{(n)} \to u$ uniformly on S_ϵ. Since the fields $X_1^{(n)}, X_2^{(n)}$ are uniformly Lipschitz continuous it follows from elementary facts about the continuous dependence of solutions of ordinary differential equations on the initial conditions (see [Hille, Theorem 3.1.1, p. 76]) that the $u^{(n)}$ are also uniformly Lipschitz continuous, so that the $y_i^{(n)}$ are uniformly bounded on S_ϵ. Since $y_i^{(n)}$ is identically 1 on the line $t_j = 0$, (3.10) implies that for sufficiently small ϵ

$$0.9 < |Y_i^{(n)}(u(t_1, t_2))|_V < 1.1$$

on S_ϵ for all n, so that by reducing ϵ, if necessary, we may assume that the $u^{(n)}$ are uniformly bi-Lipschitz on S_ϵ. From this it follows that $\kappa_i{}^{(n)}(u^{(n)}(t_1, t_2))$ tends to $\kappa_i(u(t_1, t_2))$ in $L^1(S_\epsilon)$. For sufficiently small $\epsilon > 0$, the system made up of (3.9) and its counterpart for y_2 can easily be seen to have a unique solution in $L^\infty(S_\epsilon)$. Indeed, this solution is the L^∞ limit of the sequence generated by the iteration

$$y_0 = (1, 1),\ y_{n+1} = (1, 1) - (I_1(\kappa, y_n), I_2(\kappa, y_n)). \tag{3.11}$$

Using this we can easily estimate $\|y - z\|_{L^1} = \|y_1 - z_1\|_{L^1} + \|y_2 - z_2\|_{L^1}$, where y and z are the solutions corresponding to kernels κ and η, respectively. Let M be an upper bound for the L^∞ norms of the components of κ and η. It follows immediately from (3.11) that for appropriately small $\epsilon > 0$ the L^∞ norms of the components of the y_n and z_n are all at most 2. We have

$$\|y_{1,n+1} - z_{1,n+1}\|_{L^1} = \int_{-\epsilon}^{\epsilon} \int_{-\epsilon}^{\epsilon} \int_0^{t_2} |\kappa_1 y_{1,n} y_{2,n} - \eta_1 z_{1,n} z_{2,n}| d\tau dt_1 dt_2,$$

where all the functions in the integrands are evaluated at (t_1, τ). Thus,

$$\begin{aligned}\|y_{1,n+1} - z_{1,n+1}\|_{L^1} &\leq \int_{-\epsilon}^{\epsilon}\int_{-\epsilon}^{\epsilon}\int_{-\epsilon}^{\epsilon} |\kappa_1 y_{1,n} y_{2,n} - \eta_1 z_{1,n} z_{2,n}| d\tau dt_1 dt_2 \\ &\leq 2\epsilon \int_{-\epsilon}^{\epsilon}\int_{-\epsilon}^{\epsilon} |\kappa_1 y_{1,n} y_{2,n} - \eta_1 z_{1,n} z_{2,n}| d\tau dt_1 \\ &\leq 8\epsilon \|\kappa - \eta\|_{L^1} + 2\epsilon M \int_{-\epsilon}^{\epsilon}\int_{-\epsilon}^{\epsilon} |y_{1,n} y_{2,n} - z_{1,n} z_{2,n}| d\tau dt_1 \\ &\leq 8\epsilon \|\kappa - \eta\|_{L^1} + 4\epsilon M \|y_n - z_n\|_{L^1}.\end{aligned}$$

Obviously, the same bound holds for $\|y_2 - z_2\|_{L^1}$, so that

$$\|y_{n+1} - z_{n+1}\|_{L^1} \leq 16\epsilon \|\kappa - \eta\|_{L^1} + 8\epsilon M \|y_n - z_n\|_{L^1}.$$

Since $y_0 = z_0 = (1, 1)$, it follows from this that

$$\|y_{n+1} - z_{n+1}\|_{L^1} \leq 16\epsilon \|\kappa - \eta\|_{L^1}/(1 - 8\epsilon M),$$

so that

$$\|y - z\|_{L^1} \leq 16\epsilon \|\kappa - \eta\|_{L^1}/(1 - 8\epsilon M). \tag{3.12}$$

Because, as we have explained, $\kappa^{(n)} = \kappa^{(n)}(u^{(n)}(t_1, t_2))$ tends to $\kappa = \kappa(u(t_1, t_2))$ in $L^1(S_\epsilon)$, it follows from (3.12) that $y^{(n)}$ tends in the $L^1(S_\epsilon)$ norm to the (unique) solution $\overline{y}$ in $L^\infty(S_\epsilon)$ of the system (3.9) with the original κ_i's. But then, by replacing the $\kappa^{(n)}$ by an appropriate subsequence, we can assume that for almost all fixed $T \in (-\epsilon, \epsilon)$, $y^{(n)}(T, t_2)$ and $\kappa^{(n)}(T, t_2)$ converge to $\overline{y}(T, t_2)$ and $\kappa(T, t_2)$, respectively, in $L^1(-\epsilon, \epsilon)$. Thus, for such T it follows from (3.8a) and (3.10) that

$$\overline{y}_1(T, t_2) = 1 - \int_0^{t_2} \kappa_1(T, t)\overline{y}_1(T, t)\overline{y}_2(T, t) dt,$$

for almost all $t_2 \in (-\epsilon, \epsilon)$ and analogously for $\overline{y}_2$. Finally, we must show that these $\overline{y}_i$ are our original y_i, defined by $Y_i = y_i X_i$. In other words, we have to show that the $y_i^{(n)}$ converge to the y_i. As we have seen, $u^{(n)} \to u$ and $X_k^{(n)}(u^{(n)}(t_1, t_2)) \to X_k(u(t_1, t_2))$ uniformly on S_ϵ, so that if we denote by $\theta^{(n)}$ the θ corresponding to $X_1^{(n)}$, $\theta^{(n)}(u^{(n)}(t_1, t_2))$ converges uniformly to $\theta(u(t_1, t_2))$ on S_ϵ. We have by (3.5)

$$u^{(n)}(t_1, b) - u^{(n)}(t_1, a) = \int_a^b y_2^{(n)}(t_1, \tau) F_2(u^{(n)}(t_1, \tau), \theta^{(n)}(u^{(n)}(t_1, \tau))) d\tau.$$

But, as we saw, on almost all of the lines $t_1 = T$, $y^{(n)}(T, t_2)$ tends to $\overline{y}(T, t_2)$ in $L^1(-\epsilon, \epsilon)$, so that for such T we have by letting $n \to \infty$ that

$$\begin{aligned}\int_a^b y_2(T, \tau) F_2(u(T, \tau), \theta(u(T, \tau))) d\tau &= u(T, b) - u(T, a) \\ &= \int_a^b \overline{y}_2(T, \tau) F_2(u(T, \tau), \theta(u(T, \tau))) d\tau,\end{aligned}$$

from which we conclude that $\overline{y}_2(T,t_2) = y_2(T,t_2)$ for almost all $t_2 \in (-\epsilon,\epsilon)$ and analogously for y_1. This yields the desired conclusion. □

Let U and $\overline{U}$ be (small) neighborhoods in V and $\overline{V}$, and let $f : U \to \overline{U}$ be an (m_1,m_2)-mapping. Let (u_1,u_2) and $(\overline{u}_1,\overline{u}_2)$ be corresponding local coordinates, so that f is given by $\overline{u} = f(u) = (f_1(u), f_2(u))$. We consider a characteristic coordinate mapping u of S_ϵ into U for the pair X_1, X_2 of principal direction fields. Obviously, $f \circ u$ is a characteristic coordinate mapping for the pair $\overline{X}_1, \overline{X}_2$. Without loss of generality we can assume that the y_k as well as the corresponding $\overline{y}_k$ for $f \circ u$ are all positive. Clearly, $\overline{y}_k = m_k y_k$. Let θ and $\overline{\theta}$ be the inclination functions for these pairs of fields. We derive equations satisfied by the ten functions

$$u_k,\ y_k, \lambda_k = \kappa_k y_k,\ \overline{u}_k, \theta, \overline{\theta} \tag{3.13}$$

of (t_1,t_2), $k = 1,2$. Note that by the curvature (2.9) the counterpart $\overline{\lambda}_k = \overline{\kappa}_k \overline{y}_k$ of λ_k is equal to $m_i\lambda_i/m_j$. In what follows, when we say that $\partial w/\partial t_i = w'$ for some functions w, w' defined a.e. on S_ϵ we mean that there is a function v equal to w a.e. on S_ϵ such that for almost all $T \in (-\epsilon,\epsilon)$, v is absolutely continuous on the line $t_j = T$ and $\partial v/\partial t_i = w'$ holds in the strict sense a.e. on it. In particular, the preceding lemma says that (3.7) holds in this sense.

Consider a rectangle $\alpha_k \le t_k \le \beta_k, k = 1,2$, in S_ϵ. Then since the arc length element $ds = y_1 dt_1$ (a.e. along 1-characteristics) and $dA = y_1 y_2 dt_1 dt_2$, (2.17) says that

$$\int_{\alpha_1}^{\beta_1} \kappa_1(t_1,\alpha_2) y_1(t_1,\alpha_2) dt_1 - \int_{\alpha_1}^{\beta_1} \kappa_1(t_1,\beta_2) y_1(t_1,\beta_2) dt_1 = -\int_{\alpha_2}^{\beta_2}\int_{\alpha_1}^{\beta_1} c_1 y_1 y_2 dt_1 dt_2$$

and

$$\int_{\alpha_2}^{\beta_2} \kappa_2(\alpha_1,t_2) y_2(\alpha_1,t_2) dt_2 - \int_{\alpha_2}^{\beta_2} \kappa_2(\beta_1,t_2) y_2(\beta_1,t_2) dt_2 = -\int_{\alpha_2}^{\beta_2}\int_{\alpha_1}^{\beta_1} c_2 y_1 y_2 dt_1 dt_2,$$

where

$$c_i(t_1,t_2) = c_i(u,\overline{u}) = m_j^2 \frac{m_i^2 \overline{K}(\overline{u}) - K(u)}{m_i^2 - m_j^2}.$$

Thus, the following equations hold a.e. on S_ϵ:

$$\lambda_1(t_1,t_2) = \lambda_1(t_1,0) + \int_0^{t_2} c_1(t_1,\tau) y_1(t_1,\tau) y_2(t_1,\tau) d\tau$$

and

$$\lambda_2(t_1,t_2) = \lambda_2(0,t_2) + \int_0^{t_1} c_2(\tau,t_2) y_1(\tau,t_2) y_2(\tau,t_2) d\tau$$

or in derivative form

$$\frac{\partial \lambda_i}{\partial t_j} = c_i y_1 y_2. \tag{3.14}$$

We also have that

$$\frac{\partial u}{\partial t_1} = y_1 F_1(u,\theta) \tag{3.15}$$

and since $\overline{y}_1 = m_1 y_1$

$$\frac{\partial \overline{u}}{\partial t_1} = m_1 y_1 \overline{F}_1(\overline{u}, \overline{\theta}), \tag{3.16}$$

where $F_i(u, \theta)$ is defined in (3.1).

As an immediate consequence of Lemma 3.1 we also have

$$\frac{\partial y_i}{\partial t_j} = -\lambda_i y_j \tag{3.17}$$

(in the sense explained above, of course). Finally, in light of (3.4), we have $D_1\theta = R_1(u, \theta)\kappa_1 + S_1(u, \theta)$, so that since $D_1\theta = \frac{\partial \theta}{\partial t_1}/y_1$ we conclude

$$\frac{\partial \theta}{\partial t_1} = \lambda_1 R_1(u, \theta) + y_1 S_1(u, \theta), \tag{3.18}$$

and analogously, using the fact that $\overline{\lambda}_k = m_i \lambda_i / m_j$,

$$\frac{\partial \overline{\theta}}{\partial t_1} = \frac{m_i \lambda_i}{m_j} \overline{R}_1(\overline{u}, \overline{\theta}) + \overline{y}_1 \overline{S}_1(\overline{u}, \overline{\theta}). \tag{3.19}$$

We are now able to analyze the sense in which the blow-up equations are satisfied for cps-mappings which are not necessarily C^3. (The argument to follow contains an alternate derivation of these equations based on the Gauss–Bonnet formula.) Let w be a finite valued measurable function on an open set $D \subset \mathbb{R}^2$. Then for almost all $p \in D$ it is true that for all $\eta > 0$

$$\frac{1}{\pi \delta^2} \lim_{\delta \to 0} A(\{\xi : |w(\xi) - w(p)| > \eta\} \cap \Delta(p, \delta)) = 0, \tag{3.20}$$

where A denotes 2-dimensional measure, and $\Delta(p, \delta)$ is the disk of radius δ about p. A point p for which (3.20) holds will be called a *point of approximate continuity* of w. For an orthonormal pair X_1, X_2 of Lipschitz continuous fields on U we denote by $E_i = E_i(X_1, X_2)$ the image under u of the set of points of approximate continuity of $\kappa_i \circ u$, and it is immediate that this definition is independent of the coordinate system used. It is easy to see that if $\kappa = \kappa_i$ a.e. in U and p is a point of approximate continuity of κ, then $\kappa_i(p)$ exists and is equal to $\kappa(p)$.

THEOREM 3.2. *Let $f : U \to \overline{U}$ be an (m_1, m_2)-mapping. Then for almost all $p \in U$, κ_i (as defined by (2.7)) exists on the entire j-characteristic C through p, and the restriction of κ_i to C is differentiable and satisfies the blow-up equation $D_j \kappa_i = \kappa_i^2 + c_i$ along it, where D_j is to be interpreted as arc length differentiation along C.*

Proof. It is clearly enough to establish the conclusion in $u(S_\epsilon)$ for any characteristic coordinate mapping u. For convenience let $i = 1$. There is a set $B \subset (-\epsilon, \epsilon)$ of measure 2ϵ and functions κ and y which coincide with $\kappa_1 \circ u$ and y_1 a.e. on each line $t_1 = T \in B$ and are such that y and $\lambda = \kappa y$ are absolutely continuous on each of these lines and satisfy $\frac{\partial \lambda}{\partial t_2} = c_1 y y_2$ and $\frac{\partial y}{\partial t_2} = -\lambda y_2$ in the strict sense a.e. on them. We can assume in addition that for all $T \in B$ almost all points of the 2-arc C_T corresponding to $t_1 = T$ are in E_1. Then at all points (T, t_2) at which the equations are satisfied, we have

$$\frac{\partial \kappa}{\partial t_2} = \frac{\partial(\lambda/y)}{\partial t_2} = \frac{y^2 c_1 y_2 + \lambda^2 y_2}{y^2} = (c_1 + \kappa^2) y_2,$$

or, in other words,

$$D_2(\kappa \circ u^{-1}) = c_1 + (\kappa \circ u^{-1})^2, \tag{3.21}$$

where D_2 is interpreted as arc length differentiation. Since κ is absolutely continuous (3.21) holds everywhere on C_T. It follows easily from this and the fact that almost all points of C_T are of points of approximate continuity of $\kappa \circ u^{-1}$ that in fact *all* points of C_T are points of approximate continuity of $\kappa \circ u^{-1}$(since the same equation holds on almost all nearby 2-characteristics). But then from the comment contained in the last sentence immediately preceding the statement of the theorem we conclude that (3.21) holds everywhere on C_T with $\kappa \circ u^{-1}$ replaced with κ_i, as desired. □

Theorem 3.2 has the following important corollary.

COROLLARY (COMPACTNESS PRINCIPLE). *Let U be a domain in V and let $P \subset \overline{V}$ be compact. Then the class of all (m_1, m_2)-mappings of U into $\overline{V}$ for which $f(U) \subset P$ is compact in the topology of uniform convergence of first derivatives on compact sets.*

Proof. It is enough to see that any $p \in V$ and $\overline{p} \in \overline{V}$ have (small) coordinate neighborhoods U_1 and $\overline{U}_1$ such that the (m_1, m_2)-mappings $f : U \to P$ for which $f(U_1) \subset \overline{U}_1$ have, when expressed in coordinate form, uniformly Lipschitz first derivatives on U_1. For sufficiently small U_1 Theorem 3.2 implies that κ_1 and κ_2 must be uniformly bounded and the curvature equations then say that the same must be true for $\overline{\kappa}_1$ and $\overline{\kappa}_2$. But then (3.4) and its counterpart for the $\overline{\kappa}_k$ and $\overline{\theta}$ imply that the first derivatives of θ and $\overline{\theta}$ are uniformly bounded on U_1 and $\overline{U}_1$ and in light of (3.1) and the fact that the Jacobian of f is completely determined by the X_k and $\overline{X}_k$ it follows that the first derivatives of the $f \in \mathcal{C}$ are indeed uniformly Lipschitz. □

We now examine the DeTurck–Yang initial value problem from the point of view of (3.14)–(3.19). Let C be a curve in U with Lipschitz continuous unit tangent and let $(g_1, g_2) = g : C \to \overline{U}$ have locally Lipschitz continuous derivative. We assume that the factor by which g changes arc length (when calculated with respect to the metrics in U and $\overline{U}$) is everywhere strictly between m_1 and m_2. We want to find the (m_1, m_2)-mappings of a neighborhood of C onto a neighborhood of $g(C)$ which coincide with g on C. We limit consideration to mappings which are orientation preserving with respect to the coordinate systems u and $\overline{u}$; trivial modifications cover the orientation-reversing mappings. Let T be a unit tangent field to C and let $\overline{T}$ be the corresponding unit tangent field $J_g T/|J_g T|$ to $\overline{C} = g(C)$. Let X_1, X_2 and $\overline{X}_1, \overline{X}_2$ be the fields associated with an (m_1, m_2)-extension f of g. Let ϕ denote the angle, calculated with respect to the metric of V, between X_1 and T; without loss of generality we can assume that $0 < \phi < \pi$. Let $\overline{\phi} \in (0, \pi)$ be the angle between $\overline{X}_1$ and $\overline{T}$. Then

$$m_1^2 \cos^2 \phi + m_2^2 \sin^2 \phi = |J_g T|_{\overline{V}} \quad \text{and} \quad \tan \overline{\phi} = \frac{m_2}{m_1} \tan \phi, \tag{3.22}$$

so that there are two possible choices for continuous X_1 along C, that is, two possibilities for θ corresponding to an (m_1, m_2)-mapping of a neighborhood of C onto a neighborhood of $g(C)$ and coinciding with g on C. The second equation in (3.22) means that $\overline{X}_1$ (i.e., $\overline{\theta}$) is determined once one of these θ is selected. It follows from the first of these equations that θ is a Lipschitz continuous function of arc length along C, and then from the second equation that $\overline{\theta}$ is also.

In order to proceed with the present discussion as well as to carry out some of the derivations in section 4 it is necessary to examine the relationship between the curvature of the curve C, that of its image under the (m_1, m_2)-mapping f, and the values along C of the κ_i associated with f, which by Theorem 3.2 exist a.e. on C. For

the moment we assume that J_f is differentiable (as a function of two variables) at almost all points of C. The following calculation will be valid a.e. on C. By reversing the direction of some of the vectors $X_1, X_2, \overline{X}_1, \overline{X}_2$, if necessary, we can assume that

$$T = \cos\phi X_1 + \sin\phi X_2 \quad \text{and} \quad \overline{T} = \cos\overline{\phi}\,\overline{X}_1 + \sin\overline{\phi}\,\overline{X}_2. \tag{3.23}$$

Let $N = -\sin\phi X_1 + \cos\phi X_2$ be the unit normal to C and let $\kappa = \kappa(p)$ denote the geodesic curvature of C defined by $\kappa N = D_T T$. Applying (2.4), (2.5), and (2.7) we see that a.e. on C there holds

$$\begin{aligned}\kappa N = D_T T &= D_T\phi(-\sin\phi X_1 + \cos\phi X_2) + \cos\phi D_T X_1 + \sin\phi D_T X_2 \\ &= D_T\phi N + \cos\phi(\cos\phi D_1 X_1 + \sin\phi D_2 X_1) + \sin\phi(\cos\phi D_1 X_2 + \sin\phi D_2 X_2) \\ &= D_T\phi N + \kappa_1\cos^2\phi X_2 - \kappa_2\cos\phi\sin\phi X_2 - \kappa_1\sin\phi\cos\phi X_1 + \kappa_2\sin^2\phi X_1 \\ &= D_T\phi N + (\kappa_1\cos\phi - \kappa_2\sin\phi)N,\end{aligned}$$

so that

$$\kappa_1\cos\phi - \kappa_2\sin\phi = \kappa - D_T\phi. \tag{3.24}$$

If $\overline{\kappa}$ and $\overline{N}$ are the analogous entities on $\overline{V}$, then we also have

$$\overline{\kappa}_1\cos\overline{\phi} - \overline{\kappa}_2\sin\overline{\phi} = \overline{\kappa} - D_{\overline{T}}\overline{\phi},$$

so that in light of the curvature (2.9)

$$\frac{\kappa_1}{m_2}\cos\overline{\phi} - \frac{\kappa_2}{m_1}\sin\overline{\phi} = \overline{\kappa} - D_{\overline{T}}\overline{\phi}.$$

In addition, it follows from the second equation in (3.22) that

$$\cos\overline{\phi} = \frac{m_1\cos\phi}{\sqrt{m_1^2\cos^2\phi + m_2^2\sin^2\phi}}$$

and

$$\sin\overline{\phi} = \frac{m_2\sin\phi}{\sqrt{m_1^2\cos^2\phi + m_2^2\sin^2\phi}}.$$

Since we also have

$$D_{\overline{T}}\overline{\phi} = \frac{1}{\sqrt{m_1^2\cos^2\phi + m_2^2\sin^2\phi}} D_T\overline{\phi}(f(p)),$$

it therefore follows that

$$\frac{m_1}{m_2}\kappa_1\cos\phi - \frac{m_2}{m_1}\kappa_2\sin\phi = \sqrt{m_1^2\cos^2\phi + m_2^2\sin^2\phi}\,\overline{\kappa} - D_T\tan^{-1}\left(\frac{m_2}{m_1}\tan\phi\right). \tag{3.25}$$

Finally, we point out that this holds for all curves C with Lipschitz continuous tangent, as can be seen by a simple approximation argument using Theorem 3.2.

It is now easy to cast the DeTurck–Yang initial value problem in a characteristic coordinate setting. Let C be a curve in U with Lipschitz continuous unit tangent and let $(g_1, g_2) = g : C \to \overline{U}$ have locally Lipschitz continuous derivative. We associate (a small piece) of C with the diagonal $L_\epsilon = \{(t, -t) : -\epsilon < t < \epsilon\}$ of S_ϵ via a one-to-one bi-Lipschitz function $u(t, -t)$ of L_ϵ into C having Lipschitz continuous derivative. The functions $\phi, \overline{\phi}$ and consequently $\theta, \overline{\theta}$ also are determined on C via (3.22) and then κ_1 and κ_2 are determined uniquely a.e. on C as the solution of the system (3.24), (3.25), so that in effect these six functions, as well as $u = (u_1, u_2)$ and $\overline{u} = (\overline{u}_1, \overline{u}_2)$, are determined on L_ϵ. By interchanging the roles of m_1 and m_2 and/or reversing the orientations of the corresponding X_i's as necessary, we can assume that $0 < \phi < \pi/2$, so that on the initial line $y_1, y_2 > 0$. Simple geometry implies that on the initial line

$$y_1(t, -t) = \left| \frac{du(t, -t)}{dt} \right|_V \cos\phi, \qquad y_1(t, -t) = \left| \frac{du(t, -t)}{dt} \right|_V \sin\phi.$$

Thus, all of the functions (3.13) are given on the initial line; these initial values for u_k, y_k, $\overline{u}_k$, $k = 1, 2$, and $\theta, \overline{\theta}$ are continuous; but for $\lambda_k = \kappa_k y_k$, they are merely bounded measurable functions. It is well known that for a system of equations of the form

$$\frac{\partial v}{\partial t_1} = A(v, w), \qquad \frac{\partial w}{\partial t_2} = B(v, w), \tag{3.26}$$

where $v = (v_1, \ldots, v_r)$ and $w = (w_1, \ldots, w_s)$ are functions of (t_1, t_2), and where A and B are Lipschitz continuous, the initial value problem with bounded measurable initial data $v(t, -t) = v_0(t), w(t, -t) = w_0(t)$, $|t| < \epsilon$, is locally well posed. Here the solutions are bounded measurable functions. The neighborhood of L_ϵ in which the solution is guaranteed to exist depends, for a given system (3.26), on the range of the initial functions $\{(v_0(t), w_0(t)) : -\epsilon < t < \epsilon\}$. Furthermore, if we are in the C^∞ or analytic category (i.e., A, B, and the initial data belong to one of these categories) then the solutions belong to the same category in any domain in which they exist.

The only thing one must do to complete this treatment of the DeTurck–Yang initial value problem is to show that the function $f = \overline{u} \circ u^{-1}$ which maps a neighborhood of the piece $u(L_\epsilon)$ of C onto a neighborhood of $g(u(L_\epsilon))$ is an (m_1, m_2)-mapping. One would expect such to be the case, but this has in fact been substantially obscured by the calculations used to arrive at the system. It is, however, not necessary to show directly that for a solution of this system, with initial data arising from a mapping g of C into $\overline{V}$ in the way described above, f is necessarily an (m_1, m_2)-mapping. Indeed, for C^∞ data (i.e., C and g) one can conclude this solely from the basic principles governing hyperbolic systems, as is explained fully in [Ge2, section 3]. (It is because this argument is based on polynomial approximation and the principle of permanence of functional equations for analytic functions that we have pointed out in several places that certain functions arising in the calculations were analytic.) One can conclude in general that f is an (m_1, m_2)-mapping simply by approximating the initial data by data in the C^∞ category and using the compactness principle together with the uniqueness of the solution of the initial value problem.

Theorem 3.2 tells us that a solution to a DeTurck–Yang initial value problem will exist in the entire (two-sided) domain of dependence unless the solution of one of the ordinary differential equations $D_j\kappa_i = \kappa_i^2 + c_i$ blows up along one of the characteristics along which this equation is valid; with obvious modifications, an analogous statement holds for the characteristic initial value problem (see discussion immediately following

Lemma 3.3 below). In particular, we will have such global existence if the initial values of the κ_i and the values of m_1, m_2, K, and $\overline{K}$ are such that the solutions of these equations never blow up.

We will need the following.

LEMMA 3.3. *Let $C \subset V$ be an arc with Lipschitz continuous tangent and let $p \in C$. Let κ_1 and κ_2 be any two bounded measurable functions on C. Let $\overline{p} \in \overline{V}$ and let $\overline{S}$ be any tangent vector to $\overline{V}$ at $\overline{p}$ with $|\overline{S}| \in (m_1, m_2)$. Then there is an open subarc C' containing p on which there are exactly two $f : C' \to \overline{V}$ with Lipschitz continuous derivative for which $f(p) = \overline{p}$ and $J_f(p)T = \overline{S}$ and such that along C the solutions to the corresponding DeTurck–Yang initial value problems have these κ_i as the curvatures of the corresponding curves of principal strain.*

Proof. Let $z = z(s)$, $-\epsilon < s < \epsilon$, be an arc length parametrization of a subarc of C with $z(0) = p$. If $\phi(s) = \phi(z(s))$, then (3.24) is simply the differential equation

$$\phi' = \kappa - \kappa_1 \cos\phi(s) + \kappa_2 \sin\phi(s).$$

If we add the initial condition $\phi(0) = \phi_0$, where $\phi_0 \in (0, \pi)$ is either of the solutions of

$$m_1^2 \cos^2\phi_0 + m_2^2 \sin^2\phi_0 = |\overline{S}|,$$

then there is a unique Lipschitz continuous solution of the corresponding initial value problem on some interval $(-\delta, \delta)$. Let $C' = z((-\delta, \delta))$. Then it is easy to see that there is an $f : C' \to \overline{V}$ with $f(p) = \overline{p}$ and $J_f(p)T = \overline{S}$ such that the geodesic curvature $\overline{\kappa}(s)$ at $f(z(s))$ as stipulated above is determined by (3.25), that is,

$$\overline{\kappa}(s) = \left(\frac{m_1}{m_2}\kappa_1 \cos\phi - \frac{m_2}{m_1}\kappa_2 \sin\phi + D_T \tan^{-1}\left(\frac{m_2}{m_1}\tan\phi\right)\right) / \sqrt{m_1^2\cos^2\phi + m_2^2\sin^2\phi}.$$

But since (3.24) and (3.25) uniquely define κ_1 and κ_2 once ϕ, κ, and $\overline{\kappa}$ are given, the solution of the DeTurck–Yang problem corresponding to initial mapping f (with the $X_k, \overline{X}_k$ chosen in accordance with the normalizing stipulations implicit in (3.23)) will have principal strain line curvatures coinciding along C' with the given κ_1 and κ_2. ∎

We now discuss the characteristic initial value problem for (m_1, m_2)-mappings, which is often easier to apply and more appropriate for the description of certain classes of such mappings as well as of individual ones. Let $C_k, k = 1, 2$, be curves on V with arc length parametrizations $w_k : [\alpha_k, \beta_k] \to V$, $\alpha_k < 0 < \beta_k$, such that the unit tangent vector fields $T_k(s)$ are Lipschitz continuous, $C_1 \cap C_2 = \{p\}$, where $p = w_1(0) = w_2(0)$, and $\langle T_1(0), T_2(0)\rangle = 0$. Given $\overline{p} \in \overline{V}$ and orthonormal tangent vectors $\overline{T}_1, \overline{T}_2$ to $\overline{V}$ at $\overline{p}$, the characteristic initial value problem for (m_1, m_2)-mappings consists of finding such a mapping f for which the C_k are m_k-characteristics and such that $f(p) = \overline{p}$ and $J_f T_k(0) = \overline{T}_k$. Of course, the possibility of high curvatures of the initial curves C_k in general precludes the existence of a solution even in a neighborhood of $C_1 \cup C_2$, but it is a relatively straightforward matter to see, by formulating this problem in terms of characteristic coordinates via the system (3.14)–(3.19), that it is well posed in a neighborhood of p. As with the Cauchy problem (i.e., the DeTurck–Yang problem) the key requirement is that the initial data for the ten functions (3.13) be Lipschitz continuous, which will clearly be the case for the data we have described. Here again one must make sure that the solution corresponds to an (m_1, m_2)-mapping. However, as we have seen, one can avoid the possibly cumbersome calculations implicit

in a direct verification by appealing to the theory of hyperbolic systems. Specifically, in this case the desired conclusion is a consequence of the fact that by using the blow-up (2.13) together with Lemma 3.3 we can arrange initial data for a DeTurck–Yang initial value problem along a C^∞ curve through p whose tangent at p is orthogonal to neither of the $T_k(0)$ in such a way that its solution will have the desired characteristics.

The remainder of this section deals with the generalization of HP nets discussed at the end of section 2. Specifically, we shall prove the following.

THEOREM 3.4. *Let U be a simply connected domain on a 2-manifold with constant Gaussian curvature K, and let X_1, X_2 be an orthonormal pair of Lipschitz continuous fields on U with curvatures κ_1, κ_2 defined by* (2.7). *Let $m_1, m_2 > 0$ and $\overline{K}$ be constants. Then the following are equivalent.*

(i) *For almost all $p \in U$, κ_i is a differentiable function of arc length on the entire j-characteristic through p along which it satisfies the ordinary differential equation $D_j\kappa_i = \kappa_i^2 + c_i$, where c_i is as given in* (2.14). *(Note that we are assuming only that one of the two equations in* (2.13) *is satisfied; that the other also holds will follow as a consequence.)*

(ii) *X_1, X_2 is an $(m_1, m_2, \overline{K})$-HP pair.*

(iii) *There is an (m_1, m_2)-mapping of U into a 2-manifold $\overline{V}$ with Gaussian curvature $\overline{K}$ whose principal strain fields are X_1 and X_2.*

Proof. (i)⇒(ii). Assume that the fields X_1, X_2 satisfy (i). For notational convenience we deal with the case $i = 1$. Let $u : S_\epsilon \to U$ be a characteristic coordinate mapping for these fields corresponding a small characteristic quadrilateral for which Lemma 3.1 holds; again without loss of generality we may assume that the y_i are positive. Then for $i = 1, 2$ there exist functions z_i which are equal to y_i a.e. on S_ϵ, which are absolutely continuous on almost all lines $t_j =$ constant, and satisfy (3.7) in the strict sense a.e. on them. Let T be such that the differential equation for κ_1 holds on the 2-arc corresponding to $t_1 = T$ and z_1 satisfies (3.7) a.e. on this segment. Let $\kappa(t) = \kappa_1(u(T,t))$ and $z(t) = z_1(T,t)$. Then the equations say

$$\kappa' = y_2(T,t)(c_1 + \kappa^2)$$

and

$$z' = -\kappa z y_2(T,t)$$

a.e. on $(-\epsilon, \epsilon)$. Thus,

$$\frac{d(\kappa z)}{dt} = \kappa' z + \kappa z' = z y_2(c_1 + \kappa^2) - \kappa^2 z y_2 = c_1 z y_2,$$

a.e. on $(-\epsilon, \epsilon)$, so that since κz is Lipschitz continuous on $(-\epsilon, \epsilon)$, it follows that for almost all T, $\alpha_2, \beta_2 \in (-\epsilon, \epsilon)$ with $\alpha_2 < \beta_2$ there holds

$$\kappa_1(u(T,\beta_2))y_1(T,\beta_2) - \kappa_1(u(T,\alpha_2))y_1(T,\alpha_2) = c_1 \int_{\alpha_2}^{\beta_2} y_1(T,t)y_2(T,t)dt.$$

Since $dA = y_1 y_2 dt_1 dt_2$ and $|du/dt_1| = y_1 dt_1$, integration with respect to T tells us that for almost all $\alpha_2 < \beta_2$ and any $\alpha_1 < \beta_1$ in $(-\epsilon, \epsilon)$, (2.17) holds with $i = 1$ for the characteristic quadrilateral $u([\alpha_1, \beta_1] \times [\alpha_2, \beta_2])$. Since by hypothesis κ_1 is continuous on almost all 2-characteristics, this is then true for all $\alpha_2 < \beta_2$. This shows that (ii) is true locally; that it is true globally follows by breaking large quadrilaterals into smaller ones.

(ii)⇒(iii) Let X_1, X_2 be an $(m_1, m_2, \overline{K})$-HP pair, and again let $u : S_\epsilon \to U$ be a characteristic coordinate mapping for these fields. Equation (3.14) holds since it was shown to follow from the HP-condition (2.17), and (3.15), (3.17), and (3.18) hold since they are consequences of the definitions of the u_k, y_k, λ_k, and θ. None of these equations involves any of the barred functions $\overline{u}_k, \overline{\theta}$; indeed, the only place any of these functions could enter these equations is in the c_i appearing in (3.14), and this does not happen because of our assumption that the Gaussian curvatures are constant. Uniqueness for characteristic initial value problems tells us that the only solution of the system (3.14), (3.15), (3.17), (3.18) for the $u_k, y_k, \lambda_k, \theta$ is the one associated with the given pair X_1, X_2. If we add (3.16) and (3.19) to the system and solve the corresponding characteristic initial value problem with the same initial data, we get an (m_1, m_2)-mapping of a neighborhood of $u(0, 0)$. But the X_1, X_2 so arising are still the original fields. This shows that the desired mapping exists in a neighborhood of each point of U; that it exists in all of this simply connected domain will then follow from the monodromy principle.

(iii)⇒(i). This is a special case of Theorem 3.2. □

4. Some applications.

4.1. Nonexistence of cps-mappings. We shall use the blow-up equations to show that there is no cps-mapping f of the Euclidean plane onto certain (complete, noncompact) manifolds $\overline{V}$. First of all, one notes that the solutions of the ordinary differential equation $y' = y^2$ regular at 0 are

$$y(x) = \frac{y(0)}{1 - y(0)x},$$

so that if $y(0) \neq 0$, the solution blows up to the right or left of 0 accordingly as $y(0)$ is positive or negative. From this it easily follows that if $c(x)$ is a nonnegative continuous function on $\mathbb{R}$ which is not identically 0, then the equation $y' = y^2 + c(x)$ has no solutions on all of $\mathbb{R}$.

We begin by noting that, as indicted in the introduction, there are no cps-mappings f of all of $\mathbb{R}^2$ onto itself other than the linear ones. In this case K as well as $\overline{K}$ are identically zero, so that both blow-up equations reduce to $\kappa_i' = \kappa_i^2$. From the above comments together with Theorem 3.2, for any such f it follows that $\kappa_i = 0$ a.e. on each i-characteristic, which means that all characteristics are straight lines. The linearity easily follows from this.

More interesting, perhaps, are situations in which there exist no cps-mappings of $\mathbb{R}^2$ onto $\overline{V}$ at all. In light of the interpretation of such mappings as deformations produced by the cryptocrystalline solidification of a planar lamina, this rules out the attainment of certain configurations as the result of such a process. Since $V = \mathbb{R}^2$, K is identically 0. To facilitate the discussion we assume that $m_1 < m_2$. From (2.14) we have $c_i = \frac{m_i^2 m_j^2}{m_i^2 - m_j^2}\overline{K}$, so that

$$\operatorname{sgn}(c_1) = -\operatorname{sgn}(\overline{K}) \quad \text{and} \quad \operatorname{sgn}(c_2) = \operatorname{sgn}(\overline{K}). \tag{4.1}$$

We have the following.

(1) *If $\overline{K}$ does not change sign on $\overline{V}$ and is not identically* 0, *then there are no cps-mappings $f : \mathbb{R}^2 \to \overline{V}$.* This follows immediately from the foregoing since if such an f were to exist in the case of nonnegative $\overline{K}$, for example, then by Theorem 3.2 and (4.1) there would exist a 1-characteristic with arc length parametrization $z =$

$z(s)$, $-\infty < s < \infty$, along which $c_2(z(s))$ is nonnegative but not identically 0, and along which $d\kappa_2(z(s))/ds = (\kappa_2(z(s)))^2 + c_2(z(s))$, which is impossible, as indicated in the opening paragraph of this section.

For the next case we consider $\overline{V}$ such that there is a C^∞ homeomorphism $u : \mathbb{R}^2 \to \overline{V}$ for which there are a finite number of disjoint closed disks $\overline{\Delta}_k = \overline{\Delta}(p_k, r_k), k = 1, \ldots, n$ (where $\overline{\Delta}(a, r)$ is the disk $|p - a| \le r$), such that u is an isometry on $\mathbb{R}^2 \setminus \overline{\Delta}_1 \cup \cdots \cup \overline{\Delta}_n$ and there is some $\epsilon > 0$ for which $K(u(p)) < 0$ for $r_k - \epsilon < |p - p_k| < r_k$, $1 \le k \le n$. We regard the interiors of the $u(\overline{\Delta}_k)$ as being bumps on an otherwise planar surface. One can produce such a bump by replacing a disk of radius r by the surface obtained by rotating the graph of $y = q(x)$, $0 \le x \le r$, about the y-axis, where $q \in C^\infty(\mathbb{R})$ is even and both $q''(x) > 0$ and $q'(x) < 0$ on some (r', r). These bumps have the desired negative curvature in a vicinity of the boundary circle and any number of them can be grafted into the plane, provided the corresponding closed disks are disjoint.

(2) *There exist no cps-mappings f of $\mathbb{R}^2$ onto such a "bumpy" plane $\overline{V}$.* Again assume that such an f existed. Let $z = z_k(s), k = 1, 2$, be arc length parametrizations of the characteristics through some point p_0 lying inside the preimage of one of the bumps with $z_k(0) = p_0$ and increasing s corresponding to the direction of X_k. The curvatures of the lines of principal strain are bounded since the preimage W of the union of the bumps is compact, and from the above discussion of blow-up in the planar case $|D_i\kappa_j(p)| \le 1/\mathrm{dist}(p, W)$ a.e. in $\mathbb{R}^2 \setminus W$. From this it follows by a simple compactness argument on the family of 2-characteristics that there is a 2-characteristic C which just touches ∂W but is disjoint from W (for example, by minimizing the area of the part of W to one side of C). It then follows from Theorem 3.2 that there are 2-characteristics C' arbitrarily close to C along which κ_1 exists and satisfies the corresponding blow-up equation. However, in light of the hypothesis, along such a C' sufficiently close to C, $c_1 \ge 0$ but is not identically 0, which is impossible by the comment at the end of the first paragraph of this section.

4.2. The hyperbolic plane $\mathbb{H}^2$. We begin by examining blow-up of the solutions of the ordinary differential equations to which the equations (2.13) reduce when both of the Gaussian curvatures K and $\overline{K}$ are constant. Upon writing

$$\gamma_i^2 = |c_i| = m_j^2 \left| \frac{m_i^2 \overline{K} - K}{m_i^2 - m_j^2} \right|,$$

(2.13) becomes $D_j\kappa_i = \kappa_i^2 \pm \gamma_i^2$, so we have only to look at the solutions of the elementary equations $\kappa' = \kappa^2 + \gamma^2$ and $\kappa' = \kappa^2 - \gamma^2$, $\gamma > 0$, $\kappa = \kappa(s)$. The general solution of the first is $\kappa(s) = \gamma \tan(\gamma s + C)$, so that the longest open interval in which a regular solution can exist has length π/γ. On the other hand, the solutions of $\kappa' = \kappa^2 - \gamma^2$ are of the form

$$\kappa(s) = \gamma \frac{1 + C\, e^{2\gamma s}}{1 - C\, e^{2\gamma s}}, \tag{4.2}$$

which is regular on the entire s-axis with range $(-\gamma, \gamma)$ when $C < 0$, reduces to the constant γ when $C = 0$, and has singularity at $s_0 = -(1/2\gamma) \log C$ when $C > 0$, in which case the range consists of the intervals $(-\infty, -\gamma)$ for $s > s_0$ and (γ, ∞) for $s < s_0$. In particular, the solution exists on all of $\mathbb{R}$ if and only if $|\kappa(0)| \le \gamma$.

Henceforth $V = \overline{V} = \mathbb{H}^2$, so that $K = \overline{K} = -1$. For convenience we also assume that $m_1 < m_2$, which of course constitutes no loss of generality. We have

$$c_i = \frac{m_j^2(1 - m_i^2)}{m_i^2 - m_j^2},$$

so that both of the equations (2.13) will be of the form $\kappa' = \kappa^2 - \gamma^2$ $(\gamma \geq 0)$ if and only if

$$m_1 \leq 1 \leq m_2. \tag{4.3}$$

Specifically, for such m_1, m_2 they are

$$D_j \kappa_i = \kappa_i^2 - \gamma_i^2, \text{ where } \gamma_i = \sqrt{\frac{m_j^2(m_i^2 - 1)}{m_i^2 - m_j^2}}. \tag{4.4}$$

Consider the characteristic initial value problem with initial m_k-characteristic $C_k, k = 1, 2$. Let C_k have the arc length parametrization $w = w_k(s)$, $\alpha_k < s < \beta_k$, where $0 \in (\alpha_k, \beta_k)$ and where $p = w_1(0) = w_2(0)$. As was pointed out in the discussion of this problem in section 3, we are in general guaranteed a solution only in a neighborhood of p. However, if we assume (4.3) and that the curvatures κ_k of the initial curves satisfy

$$|\kappa_k(s)| \leq \gamma_k \text{ a.e. on } (\alpha_k, \beta_k), k = 1, 2,$$

then the comment in the paragraph immediately preceding the statement of Lemma 3.3 implies that the solution exists in the entire characteristic quadrilateral determined by the C_k. Among other things this means that C_1 and C_2 are simple curves and $C_1 \cap C_2 = \{p\}$. Thus, in light of the facts that $\gamma_1^2 + \gamma_2^2 = 1$ and that γ_1 can take any value in $[0, 1]$ with appropriate m_1 and m_2 satisfying (4.3), we have established the following.

THEOREM 4.1. *Let C_1 and C_2 be curves in $\mathbb{H}^2$ whose arc length parametrizations have locally Lipschitz derivatives and which meet orthogonally at p. Let $\lambda_1, \lambda_2 > 0$ satisfy $\lambda_1^2 + \lambda_2^2 \leq 1$. If the (unsigned) geodesic curvature of C_k is bounded above by λ_k, $k = 1, 2$, then these curves are both simple and p is their only common point.*

With exactly the same hypotheses on the curves this theorem holds in the n-dimensional hyperbolic space $\mathbb{H}^n$ as well. To prove this, it suffices to show that p is the only common point when C_1 and C_2 are both simple curves. Indeed, if we have established this and C_1 and C_2 satisfy the hypotheses but C_1 is not simple, then we can replace C_2 by a geodesic E_2 which joins two points of a simple subarc E_1 of C_1 and thereby obtain a contradiction since the curvature of E_2 is everywhere 0 and that of E_1 is bounded by λ_1. Thus we shall assume that C_1 and C_2 are both simple. Assume that $n \geq 3$ and that they have a second point of intersection q. Let $w = w_k(s)$ be corresponding arc length parametrizations with $w_k(0) = p$ and $w_k(a_k) = q$, $k = 1, 2$. A simple compactness argument allows us to assume that the pair C_1, C_2 minimizes $a_1 + a_2$, i.e., the sum of the lengths of the two arcs pq. Henceforth $\operatorname{dist}(z_1, z_2)$ will denote the geodesic distance between points $z_1, z_2 \in \mathbb{H}^n$. Then $\frac{d}{ds}\operatorname{dist}(p, w_k(s)) > 0$ on $(0, a_k)$, since if it were equal to 0 for some $b \in (0, a_k)$, then C_k would be orthogonal to the geodesic joining p to $w_k(b)$, and this would give us a new pair of simple curves for which the sum of lengths of the two arcs joining the two intersection points is smaller than $a_1 + a_2$. Let $\epsilon > 0$. Then there exists a new pair of simple curves C_1'

and C_2' with C^∞ arc length parametrizations v_k on $(-1, l_k + 1)$ for which

(i) $l_k \le a_1 + a_2 + \epsilon$;
(ii) the corresponding curvatures $\kappa_k(s)$ satisfy $\kappa_k(s) \le \lambda_k + \epsilon$ on $(-1, l_k + 1)$, $k = 1, 2$;
(iii) $v_k(0) = p$, $k = 1, 2$;
(iv) $v_k(l_k) = q, k = 1, 2$;
(v) $\operatorname{dist}(p, v_k(s))$ increases on $(0, l_k)$;
(vi) C_1' and C_2' are orthogonal at their common initial point p;
(vii) for no $s \in (0, l_k]$ is the geodesic which joins p to $v_k(s)$ tangent to C_k' at $v_k(s)$.

Let $V_k(s), s \in (0, l_k]$ be the unit tangent vector at O to the geodesic ray emanating from p and passing through $v_k(s)$. It follows from (vii) that $|V_k'(s)| > 0$ on $(0, l_k]$. It is also easy to see that $\lim_{s\to 0^+} |V_k'(s)|$ exists. We claim that there exist $s_k \in (0, l_k]$ such that

$$A(s_1, s_2) = \int_0^{s_1} |V_1'(s)|\, ds + \int_0^{s_2} |V_2'(s)|\, ds = \pi/2 \tag{4.5}$$

and

$$\operatorname{dist}(p, v_1(s_1)) = \operatorname{dist}(p, v_2(s_2)). \tag{4.6}$$

To see this, consider $A(t_1, t_2)$ for $(t_1, t_2) \in Q = [0, l_1] \times [0, l_2]$. Then $A(0, 0) = 0$ and, because C_1' and C_2' are orthogonal at p, $A(l_1, l_2) \ge \pi/2$. Furthermore, since $|V_k'(s)| > 0$ on $(0, l_k]$, $A(t_1, t_2)$ is strictly increasing in each of its arguments. Thus the set $S = \{(t_1, t_2) \big| A(t_1, t_2) = \pi/2\}$ is a curve which joins the union of the left-hand side and bottom of Q to the union of its right-hand side and top. (This curve could degenerate to the point (l_1, l_2).) But (v) implies that there are increasing continuous functions $\tau_k, k = 1, 2$, which map $[0, 1]$ onto $[0, l_k]$ such that $\operatorname{dist}(p, v_1(\tau_1(t))) = \operatorname{dist}(p, v_2(\tau_2(t)))$, $t \in [0, 1]$. This means that there must be a $t \in (0, 1]$ such that $(\tau_1(t), \tau_2(t)) \in S$, so that (4.5) and (4.6) hold with $(s_1, s_2) = (\tau_1(t), \tau_2(t))$.

We consider the following mappings from a domain in $\mathbb{H}^2$ into $\mathbb{H}^n$. Let $O \in \mathbb{H}^2$ and T^* be a fixed unit vector in the tangent space of $\mathbb{H}^2$ at O. We define the continuous function T_k from the interval $[0, t_k]$ to the set of unit tangent vectors to $\mathbb{H}^2$ at O by $T(0) = T^*$ and $|T_k'(s)| = |V_k'(s)|$, where $T_k(s)$ moves in the positive sense as s increases when $k = 1$, and in the negative sense when $k = 2$. Let $G_k(s)$ be the geodesic ray emanating from O in the direction $T_k(s)$ and let $G_k(s, \sigma)$ be the point on $G_k(s)$ at distance σ from O, $0 \le s \le s_k$, $0 < \sigma$. Let F_k map the sector of $\mathbb{H}^2$ made up of the $G_k(s)$, $0 \le s \le s_k$, into $\mathbb{H}^n$ in such a way that $F_k(G_k(s, \sigma))$ is the point on the geodesic ray emanating from p through $v_k(s)$ whose distance from p is σ. One easily sees that F_k is an isometry (as a mapping between surfaces) and that it is locally one-to-one, so that F_k^{-1} is well defined. Let $\overline{C}_k$ be the preimage of C_k' under F_k. Since F_k is an isometry, the curvature of $\overline{C}_k$ at $F_k(v_k(s))$ is the curvature of C_k at $v_k(s)$ when calculated from the point of view of C_k' as a curve in the submanifold made up of the geodesics joining its points to p; this curvature is at most $\kappa_k(s)$. Thus, the curvature of $\overline{C}_k$ is bounded above by $\lambda_k + \epsilon$. Let $\overline{C}_2'$ be the curve onto which $\overline{C}_2$ is carried when $\mathbb{H}^2$ is rotated about O through a positive angle of $\pi/2$. Then from our construction $\overline{C}_1$ and $\overline{C}_2'$ are simple arcs in $\mathbb{H}^2$ which meet orthogonally at O, intersect again at their other endpoint, have lengths bounded by $a_1 + a_2 + \epsilon$, and have curvatures bounded, respectively, by $\lambda_1 + \epsilon$ and $\lambda_2 + \epsilon$. If we allow ϵ to tend to 0, then a simple compactness argument will provide curves in $\mathbb{H}^2$ which satisfy the hypotheses of Theorem 4.1 in $\mathbb{H}^2$ but not the conclusion. This contradiction proves

that the theorem is indeed true in $\mathbb{H}^n$. As an immediate consequence we obtain the following result due to Epstein [E1], [E2].

COROLLARY. *A curve in $\mathbb{H}^n$ whose curvature is everywhere bounded by 1 cannot intersect itself.*

We now give a very simple and quite explicit description of all of the cps-mappings of the entire space $\mathbb{H}^2$ into itself. Actually, it is easy to see that if $f : \mathbb{H}^2 \to \mathbb{H}^2$ is an (m_1, m_2)-mapping, then f is one-to-one and onto, so that we shall speak of the cps-self-homeomorphisms of $\mathbb{H}^2$. Fix a point $O \in \mathbb{H}^2$, and consider any (m_1, m_2)-mapping $f : \mathbb{H}^2 \to \mathbb{H}^2$, again with the nonrestrictive assumption that $m_1 < m_2$. Let C_k be the k-characteristic passing through O parametrized with respect to arc length by w_k, where $w_1(0) = w_2(0) = O$. Since all characteristics of f have infinite length in both directions, it follows from the above discussion that (4.3) holds. It furthermore follows from the initial comments that we must have $|\kappa_k| \leq \gamma_k$ a.e. on C_k, and conversely, the discussion of existence and blow-up of the preceding section shows that if these bounds are satisfied then there exists a corresponding (m_1, m_2)-mapping, which, moreover, is uniquely determined by the two functions κ_1, κ_2 once we assign the image of O and directions corresponding in the image to the tangent directions of the C_k at O. (Note that by Theorem 4.1 the conditions $|\kappa_k| \leq \gamma_k$ automatically imply that C_1 and C_2 are simple and only cross at O.) Thus we have the following.

THEOREM 4.2. *Let $O \in \mathbb{H}^2$ be fixed. There is a one-to-one correspondence between cps-self-homeomorphisms of $\mathbb{H}^2$ and 6-tuples $(m_1, m_2, C_1, C_2, \overline{O}, T)$, such that*

(i) *$m_1 \leq 1 \leq m_2$;*
(ii) *C_1 and C_2 are curves, of infinite length in both directions, with Lipschitz continuous unit tangent vectors and whose (unsigned) geodesic curvatures κ_k are bounded by the numbers γ_k defined in* (4.4)*;*
(iii) *$\overline{O} \in \mathbb{H}^2$;*
(iv) *T is an orthogonal transformation of the tangent space at O onto the tangent space at $\overline{O}$.*

For each such 6-tuple the mapping is the solution of the corresponding characteristic initial value problem.

Before continuing we point out that the blow-up conditions allow one to completely answer the following question: Given simple curves C and $\overline{C}$ on $\mathbb{H}^2$, of infinite length in both directions, and whose arc length parametrizations have locally Lipschitz continuous derivatives, give necessary and sufficient conditions on a mapping $f : C \to \overline{C}$ for which $|df/ds|$ is locally Lipschitz continuous and satisfies $m_1 < |df/ds| < m_2$ a.e. on C, such that the corresponding DeTurck–Yang initial value problems have global solutions. To do this we proceed as follows. Let $m_1 \leq 1 \leq m_2$, since otherwise there are no global (m_1, m_2)-mappings of $\mathbb{H}^2$ onto itself by Theorem 4.2. Let $z(s)$, $-\infty < s < \infty$, be an arc length parametrization of a simple curve C in V and let $\overline{z}(s) = f(z(s))$. Let $T = T(s)$ be the corresponding unit tangent vector to C at $z(s)$, $\overline{S} = \overline{S}(s) = J_f T(s)$, and $\overline{T} = \overline{S}/|\overline{S}|$. We assume that $|\overline{S}(s)|$ lies everywhere between m_1 and m_2 and shall apply the notation, normalizations, and calculations of the paragraph immediately following the proof of the corollary to Theorem 3.2 in section 3. Rewriting (3.24) and (3.25) slightly we have

$$\kappa_1 \cos\phi - \kappa_2 \sin\phi = \kappa - \phi' \tag{4.7}$$

and

$$\begin{aligned} \frac{m_1}{m_2}\kappa_1 \cos\phi - \frac{m_2}{m_1}\kappa_2 \sin\phi &= |\overline{S}|\overline{\kappa} - D_T \tan^{-1}\left(\frac{m_2}{m_1}\tan\phi\right) \\ &= |\overline{S}|\overline{\kappa} - m_1 m_2 \phi'/|\overline{S}|^2, \end{aligned} \tag{4.8}$$

so that solving for κ_1 and κ_2 we find

$$\begin{bmatrix}\kappa_1\\ \kappa_2\end{bmatrix} = \frac{1}{D}\begin{bmatrix}-\frac{m_2\sin\phi}{m_1} & \sin\phi\\ -\frac{m_1\cos\phi}{m_2} & \cos\phi\end{bmatrix}\begin{bmatrix}\kappa-\phi'\\ |\overline{S}|\overline{\kappa}-m_1m_2\phi'/|\overline{S}|^2\end{bmatrix},$$

where $D=(m_1^2-m_2^2)\frac{\sin\phi\cos\phi}{m_1m_2}$. Thus we find from our analysis of the blow-up of the κ_i that a necessary and sufficient condition for the solution of the DeTurck–Yang initial value problem to exist in all of $\mathbb{H}^2$ is that the following hold a.e. for $-\infty<s<\infty$:

$$|(\kappa-\phi')\frac{m_2\sin\phi}{m_1}-(|\overline{S}|\overline{\kappa}-m_1m_2\phi'/|\overline{S}|^2)\sin\phi|\le m_2|D|\left(\frac{(1-m_1^2)}{m_2^2-m_1^2}\right)^{\frac{1}{2}},$$

$$|(\kappa-\phi')\frac{m_1\cos\phi}{m_2}-(|\overline{S}|\overline{\kappa}-m_1m_2\phi'/|\overline{S}|^2)\cos\phi|\le m_1|D|\left(\frac{(m_2^2-1)}{m_2^2-m_1^2}\right)^{\frac{1}{2}}.$$

These bounds are, admittedly, not particularly revealing but they become considerably more so when we limit ourselves to the case in which ϕ is constant, that is, when the initial mapping of the curve C onto $\overline{C}$ has length change $|\overline{S}|=\sigma$, a constant. Since in this case we have $\phi'=0$, the conditions simplify to

$$|m_2\kappa-m_1\overline{\kappa}\sigma|\le\sqrt{(m_2^2-m_1^2)(1-m_1^2)}\,|\cos\phi|,$$

$$|m_1\kappa-m_2\overline{\kappa}\sigma|\le\sqrt{(m_2^2-m_1^2)(m_2^2-1)}\sin\phi.$$

Finally, we derive some sharp values for the radius of convexity for cps-mappings in $\mathbb{H}^2$. Returning to (4.7) and (4.8) above, we see that

$$\overline{\kappa}=\frac{1}{|\overline{S}|}\Big(\frac{m_1}{m_2}\kappa_1\cos\phi-\frac{m_2}{m_1}\kappa_2\sin\phi+m_1m_2\phi'/|\overline{S}|^2\Big),$$

so that, since $|\overline{S}|^2=m_1^2\cos^2\phi+m_2^2\sin^2\phi$, we have by (4.7) that

$$|\overline{S}|^3\overline{\kappa}=m_1m_2(\kappa_2\sin^3\phi-\kappa_1\cos^3\phi+\kappa)+\frac{m_1^3}{m_2}\kappa_1\cos^3\phi-\frac{m_2^3}{m_1}\kappa_2\sin^3\phi.$$

Thus, writing $\mu=(\frac{m_2}{m_1})^2$ we have

$$\frac{|\overline{S}|^3\overline{\kappa}}{m_1m_2}=\kappa-(\mu-1)\left(\frac{\kappa_1}{\mu}\cos^3\phi+\kappa_2\sin^3\phi\right)\text{ a.e. on }C.$$

Let $\Delta=\Delta(R,a)$ denote the disk of radius R and centered at a in $\mathbb{H}^2$ and let $f:\Delta\to\mathbb{H}^2$ be an (m_1,m_2)-mapping, which, without loss of generality we assume to be orientation preserving. We apply the above calculations to the curve $\partial\Delta$ with positive orientation so that N and $\overline{N}$ are inward pointing normals (see (3.23) and the sentence which follows it). The curve $\partial f(\Delta)$ is convex if and only if $\overline{\kappa}=\langle D_{\overline{T}}\overline{T},\overline{N}\rangle\ge 0$ a.e. on $\partial\Delta$, that is, if and only if

$$\kappa\ge(\mu-1)\left(\frac{\kappa_1}{\mu}\cos^3\phi+\kappa_2\sin^3\phi\right). \tag{4.9}$$

If p is a point of Δ at distance d from $\partial\Delta$, then it follows from (4.2) that the greatest value that $\kappa_i(p)$ can have is

$$\kappa_i^{\max} = \gamma_i \frac{e^{2\gamma_i d}+1}{e^{2\gamma_i d}-1} = \gamma_i \coth(\gamma_i d), \tag{4.10}$$

since otherwise f would have to have a singularity inside Δ. It is well known and easily calculated that the hyperbolic geodesic curvature $k(r)$ of a circle of hyperbolic radius r is given by

$$k(r) = \frac{1+\tanh^2(r/2)}{2\tanh(r/2)}.$$

It then follows from (4.9) that $f(\Delta(r,a))$ is convex provided that

$$k(r) \geq (\mu-1)\max\left\{\frac{\gamma_1}{\mu}\coth(\gamma_1(R-r)), \gamma_2\coth(\gamma_2(R-r))\right\}. \tag{4.11}$$

For fixed $m_1 \leq 1 \leq m_2$, $R>0$ the right-hand side is increasing, so that since the left-hand side is decreasing, there is a unique $\rho = \rho(m_1, m_2, R)$ for which they coincide.

THEOREM 4.3. *Let $m_1 \leq 1 \leq m_2$, $R>0$. Then $\rho(m_1,m_2,R)$ is the largest r such that all (m_1,m_2)-mappings of $\Delta(R,a)$ into $\mathbb{H}^2$ map $\Delta(r,a)$ onto simply covered convex domains.*

Proof. That the images of the concentric disk of radius $\rho(m_1,m_2,r)$ are all convex follows from the preceding discussion. Thus we have only to show that this ρ cannot be replaced by any larger number. Let i be the index corresponding to the maximum in (4.11). Let C_j be a geodesic through a and let $q \in C_j$ be at distance ρ from a. Let $d > R-\rho$, and let C_i be a curve orthogonal to C_j at q whose geodesic curvature is 0 everywhere except on a small neighborhood N of q along which it is given by the expression in (4.10), with the "concave side" of N towards the shorter of the two arcs into which q divides C_j. It is clear then that for sufficiently small N the solution f to the characteristic initial value problem for (m_1,m_2)-mappings with these characteristics exists in all of $\Delta(R,a)$. But given any $r > \rho$, for a $d > R-\rho$ sufficiently close to $R-\rho$, (4.9) will be violated for the circle centered at a and of radius r, that is, the image of the interior of this circle will not be a convex domain. ∎

5. Comments. In closing we touch on a few of the many questions about cps-mappings that naturally suggest themselves. First of all, there are reasons to believe that the Jacobian of a C^1-mapping between 2-manifolds having constant principal strains is necessarily locally Lipschitz continuous. A partial result in this direction was given in [Ge1], where it was shown that in the planar case this conclusion is valid under the stronger assumption that the derivatives of the mapping satisfy a Hölder condition with exponent $\alpha > (\sqrt{5}-1)/2$, and the arguments given there can be strengthened to extend this result to the general manifold context with the lower bound decreased to $1/2$.

In section 4 we considered only the radius of convexity problem in $\mathbb{H}^2$ under the assumption that $m_1 \leq 1 \leq m_2$ because for other values of the principal stretches there are no (m_1,m_2)-mappings of $\Delta(R,a)$ into $\mathbb{H}^2$ when R is sufficiently large. This leads us to the problem of determining the radius of the largest disk on a complete manifold of constant Gaussian curvature K on which there exist (m_1,m_2)-mappings into a manifold of constant Gaussian curvature $\overline{K}$. In light of the opening sentences

of section 1 the answer to this question, and more generally the determination of maximal domains of existence for cps-mappings on manifolds, would have an obvious bearing on the appearance of flaws in cryptocrystalline films.

Theorem 4.2 gives a complete description of all cps-mappings of $\mathbb{H}^2$ onto itself, and we have done the same [Ge4] for two planar domains (the half-plane and the exterior of a disk), but it would appear that the nonlinear hyperbolic nature of the underlying equations precludes such a description in any appreciable generality. Moreover, it is most likely that even for many "nice" domains in $\mathbb{R}^2$ there are no such mappings at all. (Although we believe this to be the case for disks, we have as yet been unable to come up with a proof.) These circumstances suggest two problems: (1) Find other manifolds for which it is possible to describe all the cps-self-homeomorphisms. (2) Find some simple conditions on a manifold which imply that this class is vacuous.

We end with a few words about cps-mappings in higher dimensions, that is, about mappings with distinct constant principal stretches between n-dimensional manifolds. The treatment of section 2 can be carried over to this more general context, but the equations that result are vastly more complicated. In the first place, the higher dimensional counterpart of the system (2.9) of curvature equations, although hyperbolic, is not diagonal, and in the second place the analogues of the blow-up equations (2.13) involve not only the principal strain line curvatures but functions that give the rate of rotation of the frames of principal strain directions as well (see [Ge2]). An example of Yin [Y] shows that there are nonaffine cps-self-homeomorphisms of $\mathbb{R}^3$, and it would be of interest to determine all such mappings. Indeed, most of the questions we have touched on in this paper can be examined in the higher dimensional context as well.

REFERENCES

[DY] D. DeTurck and D. Yang, *Existence of elastic deformations with prescribed principal strains and triply orthogonal systems*, Duke Math. J., 51 (1984), pp. 243–260.

[E1] C. Epstein, *Envelopes of Horospheres and Weingarten Surfaces in Hyperbolic* 3*-Space*, unpublished manuscript, 1985.

[E2] C. Epstein, *An Analogue of Schur's Theorem for Hyperbolic Space*, unpublished manuscript, 1985.

[Ge1] J. Gevirtz, *On planar mappings with prescribed principal strains*, Arch. Rational Mech. Anal., 117 (1992), pp. 295–320.

[Ge2] J. Gevirtz, *A diagonal hyperbolic system for mappings with prescribed principal strains*, J. Math. Anal. Appl., 176 (1993), pp. 390–403.

[Ge3] J. Gevirtz, *Hencky-Prandtl nets with isolated singularities*, Ann. Acad. Sci. Fenn. Math., 25 (2000), pp. 187–238.

[Ge4] J. Gevirtz, *Boundary Behavior and the Transformation Problem for Planar Mappings with Constant Principal Strains*, in preparation.

[He] W.S. Hemp, *Optimum Structures*, Clarendon Press, Oxford, UK, 1973.

[Hic] N. Hicks, *Notes on Differential Geometry*, Van Nostrand, Princeton, NJ, 1965.

[Hil] R. Hill, *The Mathematical Theory of Plasticity*, Clarendon Press, Oxford, UK, 1964.

[Hille] E. Hille, *Lectures on Ordinary Differential Equations*, Addison-Wesley, Reading, MA, 1969.

[L] P.D. Lax, *Development of singularities of solutions of nonlinear hyperbolic partial differential equations*, J. Math. Phys., 5 (1964), pp. 611–613.

[Y] W.-L. Yin, *Two families of finite deformations with constant strain invariants*, Mech. Res. Comm. 10 (1983), pp. 127–132.

SIAM J. MATH. ANAL.
Vol. 32, No. 4, pp. 760–777

GLOBAL NONNEGATIVE SOLUTIONS OF A NONLINEAR FOURTH-ORDER PARABOLIC EQUATION FOR QUANTUM SYSTEMS*

ANSGAR JÜNGEL† AND RENÉ PINNAU‡

Abstract. The existence of nonnegative weak solutions globally in time of a nonlinear fourth-order parabolic equation in one space dimension is shown. This equation arises in the study of interface fluctuations in spin systems and in quantum semiconductor modeling. The problem is considered on a bounded interval subject to initial and Dirichlet and Neumann boundary conditions. Further, the initial datum is assumed only to be nonnegative and to satisfy a weak integrability condition. The main difficulty of the existence proof is to ensure that the solutions stay *nonnegative* and exist *globally* in time. The first property is obtained by an exponential transformation of variables. Moreover, entropy-type estimates allow for the proof of the second property. Results concerning the regularity and long-time behavior are given. Finally, numerical experiments underlining the preservation of positivity are presented.

Key words. higher order parabolic PDE, global solution, existence, uniqueness, positivity, entropy

AMS subject classifications. 35K35, 35B99, 35G30

PII. S0036141099360269

1. Introduction.

In the last years, the study of *nonnegative* or *positive* solutions to parabolic fourth-order equations has attracted a lot of attention in the mathematical literature (see [Ber98], [BP98], [dPGG98], [Grü95], and the references therein). In particular, it was shown that certain degenerate equations of the form

$$h_t = -(f(h)h_{xxx})_x + (g(h)h_x)_x \tag{1.1}$$

allow for positive solutions if the functions f and g satisfy certain growth conditions [dPGG98]. Equation (1.1) appears in the context of surface dominated motion of thin viscous films and spreading droplets or plasticity (for an overview see [Ber98] and the references therein). When $f(h) = h,\ g(h) \equiv 0$, this equation especially arises in the modeling of droplet breakup in a Hele–Shaw cell, where the variable h describes the thickness of a neck between two masses of fluid.

Clearly, maximum principles are in general not available for fourth-order equations such that the positivity or nonnegativity property has to be proved by other techniques. The main ingredient is to exploit the special nonlinear structure of (1.1) introduced by the degenerate mobility $f(h)$, i.e., $f(h) = h^\alpha$ as $h \to 0$ for some $\alpha > 0$. This allows for nonlinear entropy dissipation, which is essential for the positivity of solutions [dPGG98].

*Received by the editors August 18, 1999; accepted for publication (in revised form) August 1, 2000; published electronically November 17, 2000. The authors acknowledge financial support from the Gerhard–Hess Programm of the Deutsche Forschungsgemeinschaft, grant number JU 359/3-1, and from the TMR Project "Asymptotic Methods in Kinetic Theory," grant ERB-FMBX-CT97-0157.

http://www.siam.org/journals/sima/32-4/36026.html

†Fachbereich Mathematik und Statistik, Universität Konstanz, D–78457 Konstanz, Germany (jungel@math.tu-berlin.de).

‡Fachbereich Mathematik, Technische Universität Darmstadt, D–64289 Darmstadt, Germany (pinnau@mathematik.tu-darmstadt.de).

In this paper we show that the fourth-order equation

$$n_t = -(n(\log(n))_{xx})_{xx} \tag{1.2a}$$

for $t > 0$ subject to the initial condition

$$n(0,x) = n_0(x) \tag{1.2b}$$

allows for nonnegative solutions.

This equation, which can be equivalently written as

$$n_t = -n_{xxxx} + \left(\frac{n_x^2}{n}\right)_{xx}, \tag{1.2c}$$

arises as a scaling limit in the study of interface fluctuations in a certain spin system [DLSS91]. The variable n describes the scaling limit of probabilities for a random variable. Problem (1.2a)–(1.2b) with periodic boundary conditions was first studied by Bleher, Lebowitz, and Speer in [BLS94]. Assuming (strictly) positive $H^1(\Omega)$-data, they showed that there exists a unique positive classical solution locally in time. For "small" initial data, the solution is even global in time. However, the problem of whether nonnegative solutions for general (nonnegative) initial data exist *globally* in time remained open. In this paper we solve this problem. Note that the equivalent formulation of (1.2a) is not degenerate such that the concept of nonlinear entropy dissipation is not applicable.

More specifically, we consider (1.2a) in the bounded domain $\Omega = (0,1)$ subject to the boundary conditions

$$n(0) = n(1) = 1, \quad n_x(0) = n_x(1) = 0. \tag{1.2d}$$

Our results extend to Dirichlet boundary conditions $n(0) \neq n(1)$, but we use (1.2d) for the sake of a smoother presentation. Although (1.2a) is (formally) derived for $\Omega = \mathbb{R}$, we study the problem in a bounded domain subject to the conditions (1.2d) for the following reason.

Equation (1.2a) also arises in the modeling of quantum semiconductor devices [Pin99a]. More precisely, the so-called quantum drift diffusion model ([PU99], [AI89]) simplifies to (1.2a) in the case of zero temperature and zero (or negligible) electric field (see also [GJ99a], [Jün98], [Jün97]). In several space dimensions the simplified and scaled equation reads

$$n_t = -2 \operatorname{div}\left(n\nabla\left(\frac{\Delta\sqrt{n}}{\sqrt{n}}\right)\right)$$

or, equivalently (assuming smooth nonvacuum solutions),

$$n_t = -\sum_{i,j} \partial_i\partial_j \left(n\,\partial_i\partial_j \log(n)\right).$$

In this context, n denotes the density of electrons in the semiconductor crystal. The expression $\Delta\sqrt{n}/\sqrt{n}$ is the so-called quantum Bohm potential. Now, in quantum semiconductor modeling, usually the boundary conditions (1.2d) are used (see [Gar94], [Pin99b]). Note that our arguments also apply to the case of periodic boundary conditions.

We show that for nonnegative initial data satisfying a certain integrability condition, there exists a generalized nonnegative solution globally in time. We stress the fact that we do not assume (strictly) positive initial data. As we impose only weak assumptions on the data, we can a priori not expect that our solutions have $L^2_{loc}(0,\infty;H^2(\Omega))$-regularity (see Proposition 3.2). On account of this fact we have to weaken our solution concept. Our main result is as follows.

THEOREM 1.1. *Assume that the initial datum n_0 is measurable and satisfies the condition*

$$\int_\Omega n_0 - \log(n_0)\,dx < +\infty. \tag{1.3}$$

Then there exists a solution n of (1.2a)–(1.2d) *satisfying*

$$n(x,t) \geq 0 \quad \text{a.e. in } (0,\infty)\times\Omega, \tag{1.4a}$$

$$n \in L^2_{loc}(0,\infty;W^{1,1}(\Omega)), \quad n_t \in L^1_{loc}(0,\infty;H^{-2}(\Omega)), \tag{1.4b}$$

$$\log(n) \in L^2_{loc}(0,\infty;H^2(\Omega)) \cap L^\infty(0,\infty;L^1(\Omega)). \tag{1.4c}$$

Further, $n(\cdot,0) = n_0$ in the sense of $H^{-2}(\Omega)$ and it holds for any $T > 0$ and any smooth test function $\phi \in C_c^\infty((0,\infty)\times\Omega)$,

$$\int_0^T \langle n_t,\phi\rangle_{H^{-2},H^2_0}\,dt = -\int_0^T\int_\Omega n\,(\log(n))_{xx}\,\phi_{xx}\,dxdt.$$

REMARK 1.1.

(a) *From condition* (1.3) *one can readily deduce that $n_0 \in L^1(\Omega)$ (see Corollary* 2.5*). Thus we only impose very weak regularity assumptions on the initial data.*

(b) *Starting from smooth, positive initial data Bleher, Lebowitz, and Speer* [BLS94] *used the stronger concept of mild solutions, employing results from semigroup theory.*

(c) *The regularity of the solutions provided by Theorem* 1.1 *is quite weak. In particular, it is not clear whether uniqueness of solutions holds in the class of functions satisfying* (1.4)*. Bleher, Lebowitz, and Speer proved in* [BLS94] *the uniqueness of solutions to* (1.2a)–(1.2b) *with periodic boundary conditions, assuming (strictly) positive $H^1(\Omega)$ initial data. These results can be easily extended to our set of boundary data.*

(d) *Considering* (1.1) *with $g(h) \equiv 0$, Bernis and Friedman* [BF90] *established a very weak solution concept, basically saying that h is a solution if for all $\phi \in L^2(0,T;H^1(\Omega))$ it holds that*

$$\int_0^T \langle h_t,\phi\rangle\,dt - \int_{\{|h|>0\}} f(h)\,h_{xxx}\,\phi_x\,dxdt = 0.$$

The proof of Theorem 1.1 is based on two ideas. The first one is to perform an exponential transformation of variables. Setting $n = e^{2u}$, (1.2a) reads in the new variable

$$\left(e^{2u}\right)_t = -2\left(e^{2u}\,u_{xx}\right)_{xx}. \tag{1.5}$$

Hence, the existence of a (generalized) solution u of (1.5) implies the existence of a nonnegative solution n of (1.2a). Exponential transformations were already successfully employed in the study of the stationary quantum hydrodynamic equations [GJ99b], [BGMS95].

Clearly, a solution $u \in L^\infty((0,\infty)\times\Omega)$ to (1.5) provides a *positive* solution n to (1.2a). However, we get only the regularity $u \in L^2_{loc}(0,\infty;L^\infty(\Omega))$ (see (1.4)) such that we can only conclude the existence of nonnegative solutions to (1.2a). This is in contrast to the stationary problem, where the positivity property immediately follows from an $H^s(\Omega)$ bound for the corresponding stationary variable u and the Sobolev embedding $H^s(\Omega) \hookrightarrow L^\infty(\Omega)$ when $s > d/2$, d being the space dimension (see [GJ99b]).

This observation motivated us to discretize (1.5) in time, which is the second main idea for the proof, yielding a sequence of elliptic problems. We show the existence of solutions $u(t_k,\cdot)$ in $H^2(\Omega)$ to the resulting elliptic problems. Hence, the approximate solutions $u(t_k,\cdot)$ are in $L^\infty(\Omega)$ and expressions like $e^{u(t_k,x)}$ are well defined.

It is worth noting that (1.2a) possesses several Lyapunov functionals which provide a priori estimates in the existence proof. It can be easily seen that the *entropy*

$$S(t) = \int_\Omega n(t)\,(\log(n(t)) - 1) + 1\,dx$$

is (formally) nonincreasing in time. This has also been observed in [BLS94]. In the case of periodic boundary conditions, also the *Fisher information*

$$\int_\Omega \left|(\sqrt{n})_x\right|^2\,dx$$

is nonincreasing in time. In addition we prove that the quantity

$$\int_\Omega n(t) - \log(n(t))\,dx$$

is nonincreasing in time. More precisely, we show that

$$\int_\Omega n(t) - \log(n(t))\,dx + \int_0^t \int_\Omega |(\log(n(t)))_{xx}|^2\,dx\,dt \le \int_\Omega n_0 - \log(n_0)\,dx$$

(in a sense to be made precise later).

For the unique solvability of the resulting elliptic systems the following monotonicity property is essential. The idea is the following: First, divide (1.2a) by $\sqrt{n}$. Then we obtain (formally)

$$(\sqrt{n})_t = A(\sqrt{n}) \stackrel{\text{def}}{=} -\frac{1}{2\sqrt{n}}(n(\log(n))_{xx})_{xx}.$$

A formal computation shows that the operator $-A(\sqrt{n})$ is monotone:

$$-\langle A(\sqrt{n_1}) - A(\sqrt{n_2}), \sqrt{n_1} - \sqrt{n_2}\rangle \ge 0 \quad \text{for suitable } \sqrt{n_1}, \sqrt{n_2}.$$

This property will be made precise in section 2. We notice that the monotonicity of the operator $-A(\sqrt{n})$ was already used in the analysis of the stationary quantum drift diffusion equations [PU99], as well as for the investigation of stability properties of the linearized transient model [Pin99a].

The paper is organized as follows. Section 2 is devoted to the proof of Theorem 1.1. In section 3 we give a result on the long-time behavior of solutions and consider regularity questions. Further, we present some numerical experiments underlining the preservation of positivity for $t > 0$.

1.1. Notation and auxiliary results. We use the standard notation for Sobolev spaces (see [Ada75]), denoting the norm of $W^{m,p}(\Omega)$ ($m \in \mathbb{N}, p \in [1,\infty]$) by $\|\cdot\|_{W^{m,p}(\Omega)}$. In the special case $p = 2$ we use $H^m(\Omega)$ instead of $W^{m,2}(\Omega)$. Further, let $H_0^m(\Omega)$ be the closure of $C_c^\infty(\Omega)$ with respect to the $H^m(\Omega)$-norm. Its dual space $(H_0^m(\Omega))^*$ is denoted by $H^{-m}(\Omega)$ and the duality pairing of $H_0^m(\Omega)$ with its dual space is given by $\langle\cdot,\cdot\rangle_{H^{-m},H_0^m}$. Moreover, for any Banach space B we define the space $L^p(0,T;B)$ with $p \in [1,\infty]$ consisting of all measurable functions $\varphi : (0,T) \to B$ for which the norm

$$\|\varphi\|_{L^p(0,T;B)} \stackrel{\text{def}}{=} \left(\int_0^T \|\varphi\|_B^p \, dt\right)^{1/p}, \quad p \in [1,\infty),$$
$$\|\varphi\|_{L^\infty(0,T;B)} \stackrel{\text{def}}{=} \sup_{t\in(0,T)} \|\varphi(t)\|_B, \quad p = \infty$$

is finite. If the time interval is clear we shortly write $\|\cdot\|_{L^p(B)}$.

In the forthcoming analysis we make frequent use of the Gagliardo–Nirenberg inequality [GT83].

LEMMA 1.2. *Let $\Omega \in \mathbb{R}$ be a bounded domain and $m \geq 1$. Furthermore, let $1 \leq p,q,r \leq \infty$, $j \in \mathbb{N}_0$ with $j < m$ and $\theta \in [j/m, 1]$ such that*

$$\frac{1}{p} = j + \theta\left(\frac{1}{r} - m\right) + (1-\theta)\frac{1}{q}$$

provided that $m - j - 1/r$ is a nonnegative integer (or else take $\theta = j/m$). Then there exists a constant $C = C(\Omega, m, j, \theta, p, q, r) > 0$ such that for all $\varphi \in W^{m,r}(\Omega) \cap L^q(\Omega)$

$$\left\|D^j\varphi\right\|_{L^p(\Omega)} \leq C\,\|\varphi\|_{W^{m,r}(\Omega)}^\theta\,\|\varphi\|_{L^q(\Omega)}^{1-\theta}.$$

2. Existence. In this section we provide the proof of Theorem 1.1, which is done in several steps. First, we introduce an exponential transformation of variables. Setting $n = e^{2u}$ we get (1.5), which will be investigated in the following. Instead of Theorem 1.1, we prove the following result, which yields Theorem 1.1 by back transforming the variables.

PROPOSITION 2.1. *Assume that the initial datum u_0 is measurable and satisfies*

$$\int_\Omega e^{2u_0} - 2u_0\,dx < \infty. \tag{2.1}$$

Then there exists a solution $u \in H_0^2(\Omega)$ of

$$\left(e^{2u}\right)_t = -2\left(e^{2u}u_{xx}\right)_{xx}$$

satisfying $e^{2u(\cdot,0)} = n_0$ in the sense of $H^{-2}(\Omega)$ and

$$u \in L^2_{loc}(0,\infty;H_0^2(\Omega)) \cap L^\infty(0,\infty;L^1(\Omega)), \tag{2.2a}$$
$$e^{2u} \in L^2_{loc}(0,\infty;W^{1,1}(\Omega)), \quad \left(e^{2u}\right)_t \in L^r_{loc}(0,\infty;H^{-2}(\Omega)) \tag{2.2b}$$

for $r \in [1, 10/9)$. Further, it holds for each $T > 0$ and each $\phi \in C_c^\infty((0,\infty)\times\Omega)$

$$\int_0^T \left\langle\left(e^{2u}\right)_t, \phi\right\rangle_{H^{-2},H_0^2} dt + 2\int_0^T\int_\Omega e^{2u}\,u_{xx}\,\phi_{xx}\,dx\,dt = 0. \tag{2.3}$$

REMARK 2.1. *Note that this result is even stronger than Theorem* 1.1*, since it implies* $n_t \in L^r_{loc}(0,\infty;H^{-2}(\Omega))$ *for* $r \in [1,10/9)$*. This gain of regularity is significantly simplifying the proof due to the reflexivity of* $L^r_{loc}(0,\infty;H^{-2}(\Omega))$ *for* $r>1$.

Second, to prove Proposition 2.1 we employ a vertical line method [Rek82], i.e., we use a semidiscretization in time. This yields a sequence of elliptic problems, for which we show that each possesses a unique positive solution.

Third, we derive a priori estimates on the sequence of approximating solutions, which allow us to perform the limit in the weak formulation. They are of energy type and of entropy type as well.

2.1. Semidiscretization. We divide the interval $[0,T]$ into N subintervals by introducing the partition $0=t_0<t_1<\cdots<t_N=T$. Setting $\tau_k \stackrel{\text{def}}{=} t_k - t_{k-1}$ we define the maximal subinterval length $\tau \stackrel{\text{def}}{=} \max_{k=1,\dots,N}\tau_k$. We assume that the partition fulfills

$$\tau \to 0 \quad \text{as } N\to\infty. \tag{2.4}$$

REMARK 2.2. *Certainly, uniform partitions satisfy* (2.4) *and are sufficient for the analytical investigations. However, as the method is also of great numerical interest as it provides a positivity preserving scheme, we allow for variable timesteps, which increases the flexibility of the method* [JP99].

For any Banach space B we define

$$PC_N(0,T;B) \stackrel{\text{def}}{=} \left\{u^{(N)}:(0,T]\to B : u^{(N)}|_{(t_{k-1},t_k]} \equiv \text{const. for } k=1,\dots,N\right\}$$

and introduce the abbreviation $u_k = u^{(N)}(t)$ for $t\in(t_{k-1},t_k]$ and $k=1,\dots,N$. Further, let $\tilde{u}^{(N)}$ denote the linear interpolant of $u^{(N)}\in PC_N(0,T;L^2(\Omega))$ given by

$$\tilde{u}^{(N)}(t,x) = \frac{t-t_{k-1}}{\tau_k}\,(u_k-u_{k-1}) + u_{k-1} \quad \text{for } x\in\Omega,\quad t\in(t_{k-1},t_k].$$

Now we discretize (1.5) in the following way.

For $k=1,\dots,N$, solve recursively the elliptic equations

$$\frac{1}{\tau_k}\left(e^{2\,u_k}-e^{2\,u_{k-1}}\right) = -2\left(e^{2\,u_k}\,u_{k,xx}\right)_{xx} \tag{2.5}$$

subject to $u_k\in H^2_0(\Omega)$ and get an approximate solution $u^{(N)}\in PC_N(0,T;H^2_0(\Omega))$. We set $e^{2\,u_0}=n_0$. Then problem (2.5) possesses a unique solution, which is the content of the following result.

PROPOSITION 2.2. *Let* u_0 *satisfy* (2.1)*. For each* $k=1,\dots,N$*, there exists a unique weak solution* $u_k\in H^2_0(\Omega)$ *of* (2.5) *fulfilling*

$$2\int_\Omega e^{2\,u_k}\,u_{k,xx}\,\phi_{xx}\,dx + \frac{1}{\tau_k}\int_\Omega e^{2\,u_k}\,\phi\,dx = \frac{1}{\tau_k}\int_\Omega e^{2\,u_{k-1}}\,\phi\,dx \tag{2.6}$$

for all $\phi\in H^2_0(\Omega)$.

For the proof of Proposition 2.2, especially for the uniqueness part, we need the following result. It states the monotonicity of the nonlinear elliptic operator, which

has already been successfully employed for the investigation of stability properties of stationary states [Pin99a].

LEMMA 2.3. *Let $u, v \in H_0^2(\Omega)$. Then the operator $A : H_0^2(\Omega) \to H^{-2}(\Omega)$ given by*

$$A(u) = -e^{-u}\left(e^{2\,u}\,u_{xx}\right)_{xx}$$

is well defined and $-A$ is monotone in the following sense:

$$-\langle A(u) - A(v), e^u - e^v\rangle_{H^{-2},H_0^2} \geq 0. \tag{2.7}$$

Proof. Since $u \in H_0^2(\Omega)$ we have by Sobolev's embedding theorems [Ada75] that $u \in C^{1,\alpha}(\bar{\Omega})$ for $0 \leq \alpha \leq 1/2$, which implies $e^u > 0$ in $\bar{\Omega}$. For $\phi \in H_0^2(\Omega)$ it holds that

$$\begin{aligned}
\langle A(u), \phi\rangle_{H^{-2},H_0^2} &= -\int_\Omega e^{2\,u}\,u_{xx}\,\left(e^{-u}\,\phi\right)_{xx}\,dx \\
&= -\int_\Omega e^{2\,u}\,u_{xx}\,\left(-u_{xx}\,\phi + u_x^2\,\phi - 2\,u_x\,\phi_x + \phi_{xx}\right)\,dx \\
&\leq \left\|e^{2\,u}\right\|_{L^\infty(\Omega)}\,\|u_{xx}\|_{L^2(\Omega)}\,\Big(\|u_{xx}\|_{L^2(\Omega)}\,\|\phi\|_{L^\infty(\Omega)} \\
&\quad + \|u_x\|_{L^4(\Omega)}^2\,\|\phi\|_{L^\infty(\Omega)} + 2\,\|u_x\|_{L^2(\Omega)}\,\|\phi_x\|_{L^\infty(\Omega)} + \|\phi_{xx}\|_{L^2(\Omega)}\Big) \\
&\leq c\left(\Omega, \|u\|_{H^2(\Omega)}\right)\,\|\phi\|_{H^2(\Omega)},
\end{aligned}$$

where we used the embedding $H^2(\Omega) \hookrightarrow W^{1,\infty}(\Omega)$ in one space dimension. Hence, A is well defined.

Now, we prove the monotonicity of A. Let $u, v \in H_0^2(\Omega)$ be given. Then $e^u \in H^2(\Omega)$ since

$$\begin{aligned}
\|(e^u)_{xx}\|_{L^2(\Omega)} &\leq \|e^u\|_{L^\infty(\Omega)}\left(\|u_{xx}\|_{L^2(\Omega)} + \|u_x\|_{L^4(\Omega)}^2\right) \\
&\leq c(\Omega)\,e^{\|u\|_{L^\infty(\Omega)}}\,\|u\|_{H^2(\Omega)}\left(1 + \|u\|_{H^2(\Omega)}\right)
\end{aligned}$$

for some positive constant $c(\Omega)$, depending only on Ω by Sobolev's embedding theorems. Now consider

$$\begin{aligned}
-\langle A(u) &- A(v), e^u - e^v\rangle_{H^{-2},H_0^2} \\
&= \left\langle\left(e^{-u}\left(e^{2\,u}\,u_{xx}\right) - e^{-v}\left(e^{2\,v}\,v_{xx}\right)\right)_{xx}, e^u - e^v\right\rangle_{H^{-2},H_0^2} \\
&= \left\langle (e^u)_{xxxx} - e^{-u}\,(e^u)_{xx}^2 - (e^v)_{xxxx} + e^{-v}\,(e^v)_{xx}^2, e^u - e^v\right\rangle_{H^{-2},H_0^2} \\
&= \int_\Omega (e^u - e^v)_{xx}^2 - (e^u)_{xx}^2 - (e^v)_{xx}^2 + e^{v-u}\,(e^u)_{xx}^2 + e^{u-v}\,(e^v)_{xx}^2\,dx \\
&= \int_\Omega \left(e^{\frac{v-u}{2}}\,(e^u)_{xx} - e^{\frac{u-v}{2}}\,(e^v)_{xx}\right)^2\,dx \\
&\geq 0,
\end{aligned}$$

which yields the assertion. ∎

Now we are in position to prove the existence and uniqueness result for the elliptic problem (2.6).

Proof of Proposition 2.2. We employ Leray–Schauder's fixed point theorem to show that there exists at least one solution. Let $k \in \{1, \dots, N\}$ be fixed and assume that $e^{2\,u_{k-1}} \in L^1(\Omega)$. Further, let $w \in H^1(\Omega)$ and $\sigma \in [0,1]$ be given and consider the following problem:
Find $u \in H_0^2(\Omega)$ with

$$2 \int_\Omega e^{2\,w}\, u_{xx}\, \phi_{xx}\, dx + \frac{\sigma}{\tau_k} \int_\Omega e^{2\,u}\, \phi\, dx = \frac{\sigma}{\tau_k} \int_\Omega e^{2\,u_{k-1}}\, \phi\, dx \tag{2.8}$$

for all $\phi \in H_0^2(\Omega)$.

On account of standard results from the theory of monotone operators [Zei90], there exists a unique weak solution $u \in H_0^2(\Omega)$ of (2.8). Thus the fixed point map $T : H^1(\Omega) \times [0,1] \to H^1(\Omega)$, given by $T(w,\sigma) = u$, is well defined.

Now let $u \in H_0^2(\Omega)$ be a fixed point of T. Then, using the test function $\phi = 1 - e^{-2\,u} \in H_0^2(\Omega)$ in (2.8) gives after integration by parts,

$$\frac{\sigma}{\tau_k} \int_\Omega e^{2\,u} - e^{2\,u_{k-1}} + e^{2(u_{k-1}-u)} - 1\, dx + 4 \int_\Omega u_{xx}^2\, dx = 8 \int_\Omega u_{xx} u_x^2\, dx.$$

With

$$\int_\Omega u_{xx} u_x^2\, dx = \frac{1}{3} \int_\Omega (u_x^3)_x\, dx = \frac{1}{3} \left(u_x(1)^3 - u_x(0)^3 \right) = 0$$

and the inequality $e^x \geq 1 + x$ for all $x \in \mathbb{R}$ we obtain

$$\frac{\sigma}{\tau_k} \int_\Omega e^{2\,u} - 2u\, dx + 4 \int_\Omega u_{xx}^2\, dx \leq \frac{\sigma}{\tau_k} \int_\Omega e^{2\,u_{k-1}} - 2u_{k-1}\, dx.$$

Therefore, using the inequality $e^x - x \geq 1$ for $x \in \mathbb{R}$ and Poincaré's inequality, there exists a constant $c > 0$ independent of u and σ such that

$$\|u\|_{H^2(\Omega)} \leq c.$$

It is easy to verify that the operator T is continuous. Hence, since the embedding $H^2(\Omega) \hookrightarrow H^1(\Omega)$ is compact, we conclude the compactness of the operator T. Furthermore, $T(w,0) = 0$ for all $w \in H^1(\Omega)$. Now the existence of at least one solution follows from Leray–Schauder's fixed point theorem.

To prove uniqueness of solutions we make use of Lemma 2.3. Assume that there exist two solutions $u, v \in H_0^2(\Omega)$ of

$$2 \left(e^{2\,u}\, u_{xx} \right)_{xx} + \frac{1}{\tau_k}\, e^{2\,u} = \frac{1}{\tau_k}\, e^{2\,u_k}.$$

Since u and v are bounded in $L^\infty(\Omega)$ we can divide the corresponding equations for u and v by e^u and e^v, respectively. Using $\phi = e^u - e^v \in H_0^2(\Omega)$ (see the proof of Lemma 2.3) as test function for the difference of the equations yields

$$\begin{aligned} -2\, \langle A(u) - A(v), e^u - e^v \rangle_{H^{-2}, H_0^2} + \frac{1}{\tau_k} \int_\Omega \left(e^u - e^v \right)^2\, dx \\ = \frac{1}{\tau_k} \int_\Omega e^{2\,u_k} \left(e^{-u} - e^{-v} \right) \left(e^u - e^v \right)\, dx \\ \leq 0, \end{aligned}$$

due to $(1/x-1/y)(x-y)\le 0$ for $x,y\in\mathbb{R}^+$. Further, the monotonicity property (2.7) implies the nonnegativity of the first term on the left-hand side. Hence, we obtain

$$\int_\Omega (e^u-e^v)^2\ dx\le 0.$$

Thus, $e^u=e^v$ in $L^2(\Omega)$ and finally $u\equiv v$, which settles the uniqueness of solutions. ∎

2.2. A priori estimates. In this section we derive a priori estimates on the sequence of approximate solutions $\big(u^{(N)}\big)_{N\in\mathbb{N}}$, which is generated by the semidiscretization. First, we show an energy type inequality and some bound on the entropy.

LEMMA 2.4. *Let u_0 satisfy (2.1). For $k=1,\ldots,N$ let $u_k\in H_0^2(\Omega)$ be the recursively defined solution of (2.6) and $u^{(N)}\in PC_N(0,T;H_0^2(\Omega))$. Then $u^{(N)}\in L^2(0,T;H_0^2(\Omega))$ and there exists a positive constant c, independent of k and N, such that*

$$\int_\Omega e^{2\,u_k}-2\,u_k\,dx+4\,\tau_k\,\|u_{k,xx}\|_{L^2(\Omega)}^2\le c, \tag{2.9a}$$

$$\sup_{t\in[0,T]}\int_\Omega e^{2\,u^{(N)}(t,x)}-2\,u^{(N)}(t,x)\,dx+4\,\Big\|u^{(N)}\Big\|_{L^2(H^2)}^2\le c. \tag{2.9b}$$

Additionally, let u_0 satisfy $\int_\Omega e^{2\,u_0}\,(2\,u_0-1)+1\,dx<+\infty$. Then it also holds that

$$\int_\Omega e^{2\,u_k}\,(2\,u_k-1)+1\,dx\le\int_\Omega e^{2\,u_{k-1}}\,(2\,u_{k-1}-1)+1\,dx, \tag{2.9c}$$

for $k=1,\ldots,N$.

Proof. Let $k\in\{1,\ldots,N\}$ be fixed and use $\phi=1-e^{-2\,u_k}$ as a test function in (2.5). Note that ϕ is an admissible test function, since $u_k\in L^\infty(\Omega)$. Integration by parts yields

$$\frac{1}{\tau_k}\int_\Omega e^{2\,u_k}-e^{2\,u_{k-1}}+e^{2(u_{k-1}-u_k)}-1\,dx=2\int_\Omega e^{2\,u_k}\,u_{k,xx}\,\big(e^{-2\,u_k}\big)_{xx}\,dx.$$

Using the well-known inequality $e^x\ge 1+x$, for $x\in\mathbb{R}$, we get

$$\frac{1}{\tau_k}\int_\Omega 2\,u_{k-1}-2\,u_k+e^{2\,u_k}-e^{2\,u_{k-1}}\,dx \le -4\int_\Omega |u_{k,xx}|^2\,dx+8\int_\Omega u_{k,xx}\,u_{k,x}^2\,dx.$$

By Young's inequality and the fact that

$$\int_\Omega u_{k,xx}\,u_{k,x}^2\,dx=\frac{1}{3}\int_\Omega\big(u_{k,x}^3\big)_x\,dx=\frac{1}{3}\big(u_{k,x}^3(1)-u_{k,x}^3(0)\big)=0$$

we derive

$$\int_\Omega e^{2\,u_k}-2\,u_k\,dx+4\,\tau_k\int_\Omega |u_{k,xx}|^2\,dx\le\int_\Omega e^{2\,u_{k-1}}-2\,u_{k-1}\,dx.$$

Thus consecutively we get

$$\int_\Omega e^{2\,u_k}-2\,u_k\,dx\le\int_\Omega e^{2\,u_{k-1}}-2\,u_{k-1}\,dx\le\cdots\le\int_\Omega e^{2\,u_0}-2\,u_0\,dx,$$

from which (2.10a) follows. Furthermore, summation with respect to k yields

$$\int_\Omega e^{2\,u_k} - 2\,u_k\,dx + 4\sum_{l=1}^{k}\tau_l\int_\Omega |u_{l,xx}|^2\,dx \le \int_\Omega e^{2\,u_0} - 2\,u_0\,dx,$$

which gives the desired estimate on $u^{(N)}$.

Now choosing $\phi = u_k$ as test function in (2.6) we obtain

$$\frac{1}{\tau_k}\int_\Omega \left(e^{2\,u_k} - e^{2\,u_{k-1}}\right) u_k\,dx = -2\int_\Omega e^{2\,u_k}\,u_{k,xx}^2\,dx. \tag{2.10}$$

Again, employing $e^x \ge 1 + x$ for $x \in \mathbb{R}$, we deduce

$$\begin{aligned}\frac{2}{\tau_k}\int_\Omega \left(e^{2\,u_k} - e^{2\,u_{k-1}}\right) u_k\,dx &= \frac{2}{\tau_k}\int_\Omega e^{2\,u_k}\left(u_k - \frac{1}{2}\right) - e^{2\,u_{k-1}}\left(u_{k-1} - \frac{1}{2}\right)dx\\ &\quad + \frac{1}{\tau_k}\int_\Omega e^{2\,u_{k-1}}\underbrace{\left(e^{2\,(u_k - u_{k-1})} - 1 - 2\,(u_k - u_{k-1})\right)}_{\ge 0}dx\\ &\ge \frac{2}{\tau_k}\int_\Omega e^{2\,u_k}\left(u_k - \frac{1}{2}\right) - e^{2\,u_{k-1}}\left(u_{k-1} - \frac{1}{2}\right)dx.\end{aligned}$$

This inequality together with (2.10) immediately implies

$$\int_\Omega e^{2\,u_k}\,(2\,u_k - 1)\,dx \le \int_\Omega e^{2\,u_{k-1}}\,(2\,u_{k-1} - 1)\,dx,$$

from which we obtain (2.9c). ∎

As a consequence of interpolation theory we derive the following estimates.

COROLLARY 2.5. *Let u_0 satisfy (2.1) and for $N \in \mathbb{N}$ let $u^{(N)} \in PC_N(0,T;H_0^2(\Omega))$ be the approximate solution. Then*

$$u^{(N)} \in L^\infty(0,T;L^1(\Omega)) \cap L^{5/2}(0,T;W^{1,\infty}(\Omega)), \quad e^{2\,u^{(N)}} \in L^{5/2}(0,T,W^{1,1}(\Omega))$$

and there exists a constant $c > 0$, independent of N, such that

$$\left\|u^{(N)}\right\|_{L^\infty(L^1)} \le c, \quad \left\|u^{(N)}\right\|_{L^{5/2}(W^{1,\infty})} \le c, \quad \left\|e^{2\,u^{(N)}}\right\|_{L^{5/2}(W^{1,1})} \le c.$$

Proof. Using Taylor's expansion we have $e^x \ge 1 + x + x^2$ for $x \ge 0$, which yields

$$\begin{aligned}\int_\Omega e^{2\,u^{(N)}} - 2\,u^{(N)}\,dx &\ge \int_\Omega e^{2\left(u^{(N)}\right)^+} - 2\left(u^{(N)}\right)^+ + 2\left(u^{(N)}\right)^-\,dx\\ &= \int_\Omega 4\left(\left(u^{(N)}\right)^+\right)^2 + 2\left(u^{(N)}\right)^- + 1\,dx,\end{aligned}$$

where $u^+ = \max(0,u)$ and $u^- = -\min(0,u)$. This estimate immediately implies $(u^{(N)})^-(t) \in L^1(\Omega)$ and $(u^{(N)})^+(t) \in L^2(\Omega)$ for any $t > 0$. Thus $u^{(N)}(t) \in L^1(\Omega)$ and it holds

$$\left\|u^{(N)}\right\|_{L^\infty(L^1)} \le c,$$

where $c = c(\Omega, u_0) > 0$ is independent of N.

To show the second inequality we use Lemma 1.2 with $m = r = 2$, $j = 1$, $p = \infty$, and $q = 1$, yielding

$$\left\|u_x^{(N)}\right\|_{L^\infty(\Omega)} \le c \left\|u^{(N)}\right\|_{H^2(\Omega)}^{4/5} \left\|u^{(N)}\right\|_{L^1(\Omega)}^{1/5},$$

which gives

$$\left\|u_x^{(N)}\right\|_{L^{5/2}(L^\infty)} \le c \left\|u^{(N)}\right\|_{L^2(H^2)}^{4/5} \left\|u^{(N)}\right\|_{L^\infty(L^1)}^{1/5}.$$

Now the assertion follows from Lemma 2.4 and the previous inequality.

From Lemma 2.4 and the first inequality we get

$$\left\|e^{2\,u^{(N)}}\right\|_{L^\infty(L^1)} \le c.$$

Hence,

$$\left\|\left(e^{2\,u^{(N)}}\right)_x\right\|_{L^{5/2}(L^1)} \le \left\|e^{2\,u^{(N)}}\right\|_{L^\infty(L^1)} \|u_x\|_{L^{5/2}(L^\infty)}$$

and the third inequality follows from the second one. ∎

The next two lemmas provide the estimates, which are necessary for the compactness arguments.

LEMMA 2.6. *Let u_0 satisfy (2.1) and for $N \in \mathbb{N}$ let $u^{(N)} \in PC_N(0,T;H_0^2(\Omega))$ be the approximate solution. Choose $p \in (1,4/3)$ and fix $q \in (2,5/2)$ such that $1/q = 2(2-1/p)/5$. Then $e^{2\,u^{(N)}} \in L^q(0,T;W^{1,p}(\Omega))$ and there exists a constant $c > 0$, independent of N, such that*

$$\left\|e^{2\,u^{(N)}}\right\|_{L^q(W^{1,p})} \le c.$$

Proof. Again, employing Lemma 1.2 we derive

$$\left\|e^{2\,u^{(N)}}\right\|_{L^p(\Omega)} \le c \left\|e^{2\,u^{(N)}}\right\|_{W^{1,1}(\Omega)}^{\frac{p-1}{p}} \left\|e^{2\,u^{(N)}}\right\|_{L^1(\Omega)}^{\frac{1}{p}}.$$

Furthermore, it holds that

$$\frac{1}{q} = \frac{2\,(p-1)}{5\,p} + \frac{2}{5}$$

and thus Hölder's inequality implies

$$\left\|\left(e^{2\,u^{(N)}}\right)_x\right\|_{L^q(L^p)} \le c \left\|e^{2\,u^{(N)}}\right\|_{L^\infty(L^1)}^{\frac{1}{p}} \left\|e^{2\,u^{(N)}}\right\|_{L^{5/2}(W^{1,1})}^{\frac{p-1}{p}} \left\|u_x^{(N)}\right\|_{L^{5/2}(L^\infty)}.$$

Hence, the assertion follows from Corollary 2.5 and (2.9c). ∎

As we want to employ compactness results, we also need some regularity on the time derivative. We introduce the linear interpolant of $e^{2\,u^{(N)}} \in PC_N(0,T,L^2(\Omega))$, defined by

$$\tilde{e}^{(N)}(t,x) \stackrel{\text{def}}{=} \frac{t-t_{k-1}}{\tau_k}\left(e^{2\,u_k(x)} - e^{2\,u_{k-1}(x)}\right) + e^{2\,u_{k-1}(x)}, \quad x \in \Omega, \quad t \in (t_{k-1}, t_k].$$

LEMMA 2.7. *Let the assumptions of Lemma* 2.6 *hold. Choose* $r \in (1, 10/9)$ *such that* $1/r = 1/2 + 1/q$. *Then*

$$\tilde{e}_t^{(N)} \in L^r(0,T;H^{-2}(\Omega)), \quad \tilde{e}^{(N)} \in L^q(0,T;W^{1,p}(\Omega))$$

and there exists a constant $c > 0$, *independent of* N, *such that*

$$\left\|\tilde{e}_t^{(N)}\right\|_{L^r(H^{-2})} + \left\|\tilde{e}^{(N)}\right\|_{L^q(W^{1,p})} \le c.$$

Proof. We introduce the solution operator $\Phi : H^{-2}(\Omega) \to H_0^2(\Omega)$, $f \mapsto \Phi[f]$ by $\Phi[f]_{xxxx} = f$. Then $\|\Phi[f]_{xx}\|_{L^2(\Omega)}$ defines a norm on $H^{-2}(\Omega)$ [Zei90]. Further, $\Phi\left[\frac{e^{2u_k} - e^{2u_{k-1}}}{\tau_k}\right]$ is an appropriate test function in (2.6), which yields

$$\int_\Omega \frac{e^{2u_k} - e^{2u_{k-1}}}{\tau_k} \Phi\left[\frac{e^{2u_k} - e^{2u_{k-1}}}{\tau_k}\right] dx = -2\int_\Omega e^{2u_k} u_{k,xx} \Phi\left[\frac{e^{2u_k} - e^{2u_{k-1}}}{\tau_k}\right]_{xx} dx$$

and using integration by parts

$$\int_\Omega \left|\Phi\left[\frac{e^{2u_k} - e^{2u_{k-1}}}{\tau_k}\right]_{xx}\right|^2 dx \le 2\left\|e^{2u_k}\right\|_{L^\infty(\Omega)} \|u_{k,xx}\|_{L^2(\Omega)} \left\|\Phi\left[\frac{e^{2u_k} - e^{2u_{k-1}}}{\tau_k}\right]_{xx}\right\|_{L^2(\Omega)}.$$

Thus we finally can estimate

$$\left\|\Phi\left[\frac{e^{2u_k} - e^{2u_{k-1}}}{\tau_k}\right]_{xx}\right\|_{L^2(\Omega)} \le 2\left\|e^{2u_k}\right\|_{L^\infty(\Omega)} \|u_k\|_{H^2(\Omega)}.$$

Now we deduce from Hölder's inequality

$$\begin{aligned}\left\|\tilde{e}_t^{(N)}\right\|_{L^r(H^{-2})}^r &= \sum_{k=1}^N \tau_k \left\|\frac{e^{2u_k} - e^{2u_{k-1}}}{\tau_k}\right\|_{H^{-2}(\Omega)}^r \\ &\le 2^r \sum_{k=1}^N \tau_k \left\|e^{2u_k}\right\|_{L^\infty(\Omega)}^r \|u_k\|_{H^2(\Omega)}^r \\ &\le 2^r \left\|e^{2u^{(N)}}\right\|_{L^q(L^\infty)}^r \left\|u^{(N)}\right\|_{L^2(H^2)}^r,\end{aligned}$$

from which we easily get the uniform boundedness of $\tilde{e}_t^{(N)}$ in $L^r(0,T;H^{-2}(\Omega))$ by Lemma 2.6 together with the embedding $W^{1,p}(\Omega) \hookrightarrow L^\infty(\Omega)$ and Lemma 2.4.

Further, we note that for arbitrary $t \in (t_{k-1}, t_k]$ it holds

$$0 \le \frac{t - t_{k-1}}{\tau_k} \le 1$$

such that (see Lemma 2.6)

$$\begin{aligned}\left\|\tilde{e}^{(N)}(t)\right\|_{W^{1,p}(\Omega)} &\le \left(1 - \frac{t - t_{k-1}}{\tau_k}\right)\left\|e^{2u_{k-1}}\right\|_{W^{1,p}(\Omega)} + \frac{t - t_{k-1}}{\tau_k}\left\|e^{2u_k}\right\|_{W^{1,p}(\Omega)} \\ &\le c.\end{aligned}$$

Thus we obtain

$$\int_0^T \left\|\tilde{e}^{(N)}(t)\right\|_{W^{1,p}(\Omega)}^q \, dt \le T\, c^q,$$

finishing the proof. ∎

2.3. Proof of the existence result. Now we are in the position to prove Proposition 2.1.

Proof. We choose a sequence of partitions of $[0,T]$ satisfying (2.4). Taking into account Lemma 2.4, it follows immediately that the sequence $\left(u^{(N)}\right)_{N\in\mathbb{N}}$ is bounded in $L^2(0,T;H^2(\Omega))$. Thus there exists a subsequence, again denoted by $\left(u^{(N)}\right)_{N\in\mathbb{N}}$, such that

$$u^{(N)} \rightharpoonup u \quad \text{weakly in } L^2(0,T;H^2(\Omega)) \text{ as } N\to\infty.$$

Furthermore, from Lemmas 2.6 and 2.7 we deduce the boundedness of $\left(\tilde{e}^{(N)}\right)_N$ in $L^q(0,T;W^{1,p}(\Omega))\cap W^{1,r}(0,T;H^{-2}(\Omega))$, where p,q,r are specified therein. Since the embedding $W^{1,p}(\Omega)\hookrightarrow L^\infty(\Omega)$ is compact for $p\in(1,4/3)$, it follows from Aubin's lemma [Sim87] that

$$L^q(0,T;W^{1,p}(\Omega))\cap W^{1,r}(0,T;H^{-2}(\Omega)) \hookrightarrow L^q(0,T;L^\infty(\Omega)) \quad \text{compactly.}$$

Hence, there exists a subsequence, not relabeled, such that

$$\tilde{e}^{(N)} \to \rho \quad \text{strongly in } L^q(0,T;L^\infty(\Omega)) \text{ for } N\to\infty.$$

As $q>2$ it also holds that $\tilde{e}^{(N)}\to\rho$ in $L^2(0,T;L^2(\Omega))$ for $N\to\infty$. Note that $\tilde{e}^{(N)}\to\rho$ in $L^2(0,T;L^2(\Omega))$ implies $e^{2\,u^{(N)}}\to\rho$ in $L^2(0,T;L^2(\Omega))$ as $N\to\infty$ (see [Rek82, p. 205]). Due to the monotonicity of the exponential function we have for all $v\in L^\infty((0,\infty)\times\Omega)$

$$\int_\Omega \left(e^{2\,u^{(N)}} - e^{2\,v}\right)\left(u^{(N)} - v\right)\, dx \ge 0.$$

The derived convergence properties are by far sufficient to pass to the limit in this inequality, which yields

$$\int_\Omega \left(\rho - e^{2\,v}\right)(u-v)\, dx \ge 0$$

for all $v\in L^\infty((0,\infty)\times\Omega)$. Again, the monotonicity of the exponential implies $\rho = e^{2\,u}$.

After this identification we can perform the limit in the weak formulation, which reads

$$\int_0^T \left\langle \tilde{e}_t^{(N)}, \phi\right\rangle_{H^{-2},H_0^2} dt = -2\int_0^T\int_\Omega e^{2\,u^{(N)}}\, u_{xx}^{(N)}\, \phi_{xx}\, dx\, dt \tag{2.11}$$

for all $\phi\in L^{r'}(0,T;H_0^2(\Omega))$ with $1/r+1/r'=1$. One easily verifies that the following convergence properties are sufficient to pass to the limit in (2.11):

$$\begin{aligned}
\tilde{e}_t^{(N)} &\rightharpoonup \left(e^{2\,u}\right)_t \quad && \text{weakly in } L^r(0,T;H^{-2}(\Omega)),\\
e^{2\,u^{(N)}} &\to e^{2\,u} && \text{strongly in } L^q(0,T;L^\infty(\Omega)),\\
u_{xx}^{(N)} &\rightharpoonup u_{xx} && \text{weakly in } L^2(0,T;L^2(\Omega)),
\end{aligned}$$

as $N\to\infty$, which proves our main result. ∎

3. Additional results and discussion. In this section we present some additional results, concerning the long-time behavior of solutions and regularity properties. Further, we give some numerical examples.

3.1. Long-time behavior. The next result states that the solution converges for $t \to \infty$ to the stationary state $n_\infty \equiv 1$ in some weak sense.

PROPOSITION 3.1. *Assume* (1.3) *and let* $n \in L^2_{loc}(0,\infty,W^{1,1}(\Omega))$ *be a solution to* (1.2). *Then it holds*

$$\|\log(n(t))\|_{L^2(\Omega)} \to 0 \quad \text{as } t \to \infty.$$

Proof. Let $n = e^{2u}$. From the proof of Lemma 2.4 it follows that

$$\int_0^\infty \int_\Omega u_{xx}^2 \, dx dt \le \int_\Omega e^{2u_0} - 2u_0 \, dx.$$

Therefore, using Poincaré's inequality, there exists a sequence $(t_m)_{m\in\mathbb{N}}$ with $t_m \to \infty$ such that

$$\|u(t_m)\|_{H^2(\Omega)} \to 0 \quad \text{as } t_m \to \infty. \tag{3.1}$$

We introduce the new entropy

$$E(t) = \int_\Omega e^{2u(t)} - 2u(t) - 1 \, dx.$$

The proof of Lemma 2.4 shows that $E(t)$ is nonincreasing:

$$E(t) \le E(s) \qquad \text{for } 0 \le s \le t < \infty.$$

The result (3.1) implies $u(t_m) \to 0$ in $L^\infty(\Omega)$, by Sobolev's embedding. Hence

$$0 \le E(t_m) \to 0 \qquad \text{as } t_m \to \infty.$$

Since E is nonincreasing, $E(t) \to 0$ for all sequences $t \to \infty$. The proof of Corollary 2.5 shows that

$$E(t) \ge \int_\Omega 4u^+(t)^2 - 2u^-(t) \, dx,$$

and thus,

$$\|u(t)\|_{L^1(\Omega)} \to 0 \qquad \text{as } t \to \infty,$$

which proves the proposition. □

3.2. Regularity. Now we investigate the regularity of solutions.

PROPOSITION 3.2. *Assume* (1.3). *If it holds*

$$\int_\Omega n_0 (\log(n_0) - 1) + 1 \, dx < +\infty,$$

then any solution $n \in L^2_{loc}(0,\infty;W^{1,1}(\Omega))$ *to* (1.2) *with*

$$\log(n) \in L^2_{loc}(0,\infty;H^2_0(\Omega)) \cap L^\infty(0,\infty;L^2(\Omega))$$

even fulfills

$$n \in L^{16/15}_{loc}(0,\infty;H^2(\Omega)).$$

REMARK 3.1. *Notice that Theorem* 1.1 *and Proposition* 3.1 *ensure the existence of a solution to* (1.2) *with the desired regularity properties.*

Proof. The proof is an easy consequence of the results derived so far combined with the Gagliardo–Nirenberg inequality. Let $n = e^{2u}$. In the following c denotes positive, but not necessarily identical, constants. We estimate

$$\begin{aligned}\left\|\left(e^{2u}\right)_{xx}\right\|_{L^2(\Omega)} &\le c\left(\left\|e^{2u}\right\|_{L^\infty(\Omega)}\|u\|_{H^2(\Omega)} + \left\|e^{2u}u_x^2\right\|_{L^2(\Omega)}\right)\\ &= c\left(\left\|e^{2u}\right\|_{L^\infty(\Omega)}\|u\|_{H^2(\Omega)} + \|(e^u)_x\|^2_{L^4(\Omega)}\right).\end{aligned}$$

Due to $u \in L^2_{loc}(0,\infty;H^2_0(\Omega)) \cap L^\infty(0,\infty;L^2(\Omega))$ and $e^{2u} \in L^2_{loc}(0,\infty;W^{1,1}(\Omega))$ it holds that (compare Lemma 2.5)

$$\left\|e^{2u}\right\|_{L^{8/3}(L^\infty)} \le c \quad \text{and} \quad \|u\|_{L^{8/3}(W^{1,\infty})} \le c.$$

Further, we have $e^u \in L^{4/3}_{loc}(0,\infty;H^2(\Omega))$, since

$$\|(e^u)_{xx}\|_{L^2(\Omega)} \le \|e^u\|_{L^\infty(\Omega)}\|u_{xx}\|_{L^2(\Omega)} + \|e^u\|_{L^2(\Omega)}\|u_x\|^2_{L^\infty(\Omega)}$$

and by multiple use of Hölder's inequality

$$\|(e^u)_{xx}\|_{L^{4/3}(L^2)} \le \|e^u\|_{L^4(L^\infty)}\|u_{xx}\|_{L^2(L^2)} + \|e^u\|_{L^\infty(L^2)}\|u_x\|^2_{L^{8/3}(L^\infty)},$$

which is finite. Now we deduce by

$$\|e^u\|_{W^{1,4}(\Omega)} \le c\,\|e^u\|^{3/8}_{L^2(\Omega)}\|e^u\|^{5/8}_{H^2(\Omega)}$$

that $e^u \in L^{32/15}_{loc}(0,\infty;W^{1,4}(\Omega))$. Hence, we get finally

$$\left\|e^{2u}\right\|_{L^{16/15}(H^2)} \le c\left(\left\|e^{2u}\right\|_{L^{16/7}(L^\infty)}\|u\|_{L^2(H^2)} + \|e^u\|^2_{L^{32/15}(W^{1,4})}\right). \qquad \square$$

3.3. Numerical examples. After the analytical discussion of problem (1.2) we present some numerical results that do not only underline the preservation of nonnegativity by the solution n. They are also indicating that the solution is *positive* for $t > 0$, even for initial data, which vanishes at some point $x_0 \in \Omega = (0,1)$. This behavior was already pointed out in [BLS94] but only for strictly positive initial data.

For the numerical experiments we choose the initial datum

$$n_0(x) = \cos^{2m}(\pi x), \quad x \in (0,1), \tag{3.2}$$

with $m = 1$ or 8 and which is compatible with the boundary data. Note that n_0 vanishes at $x_0 = 1/2$ such that $\log(n_0)$ has a singularity there. However, it still holds that $\log(n_0) \in L^1(\Omega)$.

For the computations we employ a fully implicit discretization of (1.2c) with a uniform time step $\tau = 10^{-8}$. Moreover, we choose a uniform space discretization

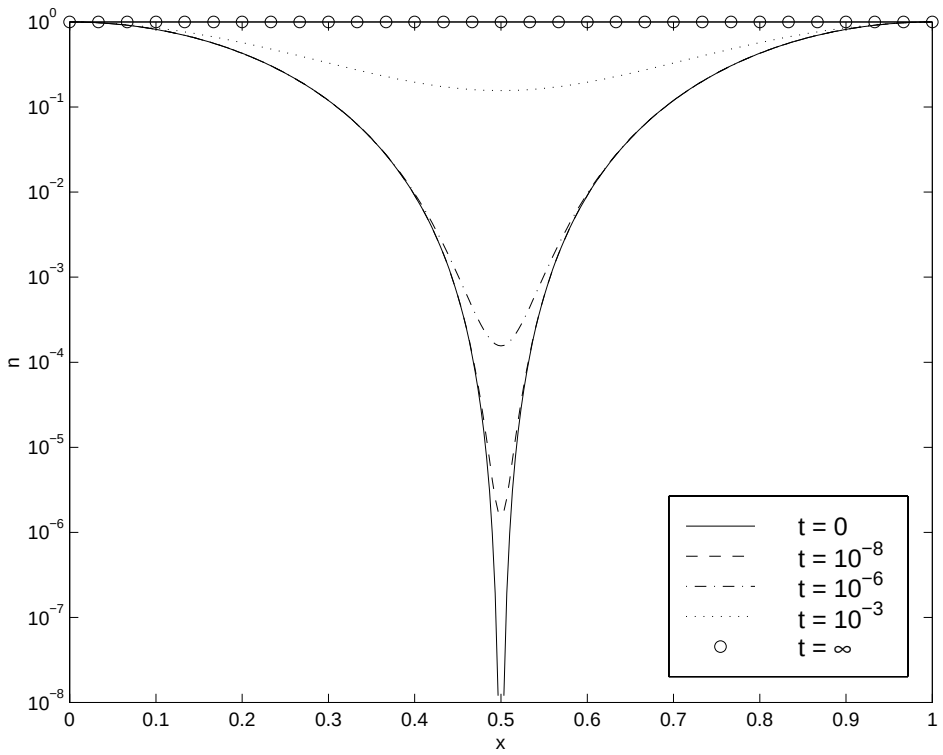


FIG. 3.1. *Evolution for* $m = 1$.

$x_l = l/M$, $l \in \{0, \dots, M\}$, $M = 300$, such that $x = 1/2$ is included in the set of nodes. Let D^+, D^- denote the standard forward, backward difference operators on this grid, respectively. Then for $l \in \{0, \dots, M\}$ and $k \in \{1, \dots, M\}$ the discrete version of (1.2c) reads

$$n_0^l = n_0(x_l),$$
$$\frac{n_k^l - n_{k-1}^l}{\tau} + D^+D^-D^+D^-n_k^l - D^+D^-\left(\frac{\left(D^+n_k^l\right)^2}{n_k^l}\right) = 0,$$

where the boundary data can be eliminated in a standard manner. These nonlinear systems are solved on each time level by a Newton-iteration, where the initial guess is chosen as the solution on the previous time level. This iteration proves to be very robust such that no damping is necessary.

REMARK 3.2. *Note that due to the fully implicit scheme, we are able to allow also for vanishing initial data. In a forthcoming paper* [JP99] *the authors investigate this approach for the so-called quantum drift diffusion model in the multidimensional case with positive initial data.*

Figure 3.1 shows the evolution of the initial datum with $m = 1$. Note that we use a logarithmic scale for the ordinate such that we cut-off all values less than 10^{-8}. Here the solution moves very fast away from zero and converges monotonically to the stationary state $n_\infty \equiv 1$. To contrast this behavior we refer to Figure 3.2, which shows the evolution for $m = 8$. Starting with *one* higher order extremum the

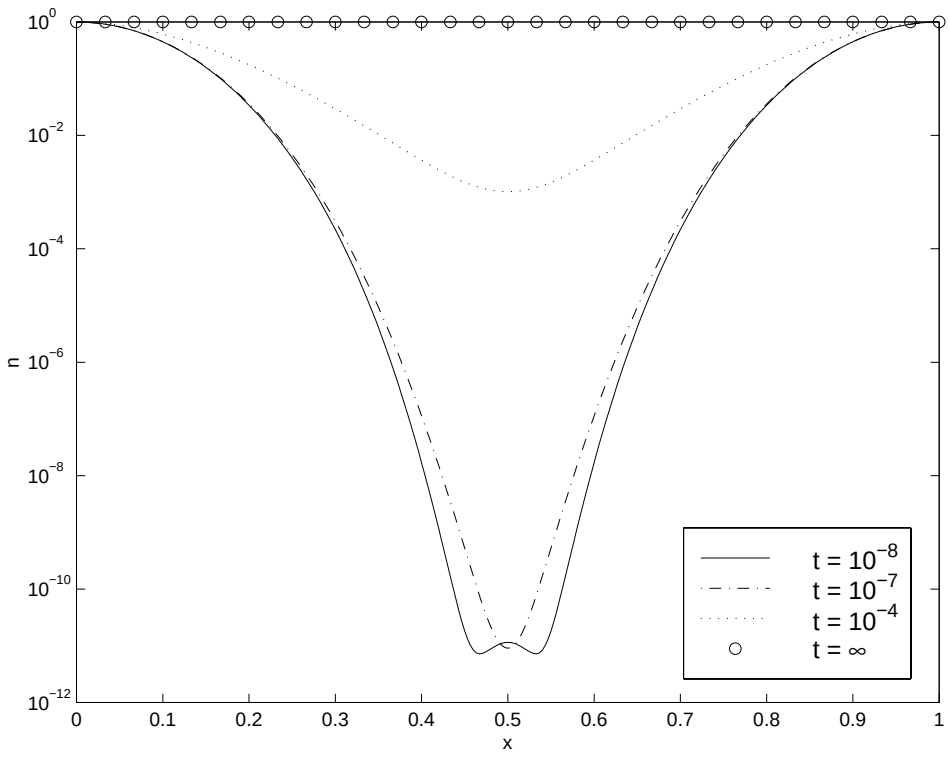


FIG. 3.2. *Evolution for* $m = 8$.

minimum bifurcates and reduces to one extremum again. We emphasize that also in this case the solution stays strictly positive for $t > 0$, although the evolution is not monotone anymore. Again, analogous results are reported in [BLS94] for strictly positive initial data.

REFERENCES

[Ada75] R. A. ADAMS, *Sobolev Spaces*, Academic Press, 1st ed., New York, 1975.

[AI89] M. G. ANCONA AND G. J. IAFRATE, *Quantum correction of the equation of state of an electron gas in a semiconductor*, Phys. Rev. B, 39 (1989), pp. 9536–9540.

[Ber98] A. L. BERTOZZI, *The mathematics of moving contact lines in thin liquid films*, Notices Amer. Math. Soc., 45 (1998), pp. 689–697.

[BF90] F. BERNIS AND A. FRIEDMAN, *Higher order nonlinear degenerate parabolic equations*, J. Differential Equations, 83 (1990), pp. 179–206.

[BGMS95] F. BREZZI, I. GASSER, P. A. MARKOWICH, AND CH. SCHMEISER, *Thermal equilibrium states of the quantum hydrodynamic model for semiconductors in one dimension*, Appl. Math. Lett., 8 (1995), pp. 47–52.

[BLS94] P. M. BLEHER, J. L. LEBOWITZ, AND E. R. SPEER, *Existence and positivity of solutions of a fourth-order nonlinear PDE describing interface fluctuations*, Comm. Pure Appl. Math., 47 (1994), pp. 923–942.

[BP98] A. L. BERTOZZI AND M. C. PUGH, *Long–wave instabilities and saturation in thin film equations*, Comm. Pure Appl. Math., 51 (1998), pp. 625–661.

[DLSS91] B. DERRIDA, J. L. LEBOWITZ, E. SPEER, AND H. SPOHN, *Fluctuations of a stationary nonequilibrium interface*, Phys. Rev. Lett., 67 (1991), pp. 165–168.

[dPGG98] R. DAL PASSO, H. GARCKE, AND G. GRÜN, *On a fourth-order degenerate parabolic equation: Global entropy estimates, existence, and qualitative behavior of solu-*

tions, SIAM J. Math. Anal., 29 (1998), pp. 321–342.

[Gar94] C. L. Gardner, *The quantum hydrodynamic model for semiconductor devices*, SIAM J. Appl. Math., 54 (1994), pp. 409–427.

[GJ99a] I. Gamba and A. Jüngel, *Positive solutions to singular second and third order differential equations for quantum fluids*, Arch. Rational Mech. Anal., to appear.

[GJ99b] M. T. Gyi and A. Jüngel, *A quantum regularization of the one–dimensional hydrodynamic model for semiconductors*, Adv. Differential Equations, 5 (2000), pp. 773–800.

[Grü95] G. Grün, *Degenerate parabolic differential equations of fourth order and a plasticity model with non–local hardening*, Z. Anal. Anwendungen, 14 (1995), pp. 541–574.

[GT83] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 1st ed., Springer-Verlag, Berlin, 1983.

[JP99] A. Jüngel and R. Pinnau, *A positivity-preserving numerical scheme for a nonlinear fourth order parabolic system*, SIAM J. Numer. Anal, submitted.

[Jün97] A. Jüngel, *A note on current–voltage characteristics from the quantum hydrodynamic equations for semiconductors*, Appl. Math. Lett., 10 (1997), pp. 29–34.

[Jün98] A. Jüngel, *A steady–state potential flow Euler–Poisson system for charged quantum fluids*, Comm. Math. Phys., 194 (1998), pp. 463–479.

[Pin99a] R. Pinnau, *The linearized transient quantum drift diffusion model—stability of stationary states*, ZAMM Z. Angew. Math. Mech., 80 (2000), pp. 327–344.

[Pin99b] R. Pinnau, *A note on boundary conditions for quantum hydrodynamic models*, Appl. Math. Lett., 12 (1999), pp. 77–82.

[PU99] R. Pinnau and A. Unterreiter, *The stationary current–voltage characteristics of the quantum drift-diffusion model*, SIAM J. Numer. Anal., 37 (1999), pp. 211–245.

[Rek82] K. Rektorys, *The Method of Discretization in Time and Partial Differential Equations*, 1st ed., Dordrecht, Boston, 1982.

[Sim87] J. Simon, *Compact sets in the space $L^p(0,T;B)$*, Ann. Math. Pura Appl. (4), 146 (1987), pp. 65–96.

[Zei90] E. Zeidler, *Nonlinear Functional Analysis and its Applications*, 1st ed., Volume II/A and II/B, Springer-Verlag, Berlin, 1990.

SIAM J. MATH. ANAL.
Vol. 32, No. 4, pp. 778–800

A FREE BOUNDARY PROBLEM ARISING IN A MODEL OF WOUND HEALING*

XINFU CHEN† AND AVNER FRIEDMAN‡

Abstract. In this paper we consider a system of two semilinear parabolic reaction-diffusion equations with a free boundary, which arises in a model of corneal epithelial wound healing. We prove that the initial-boundary value problem has a unique solution and that complete healing is achieved in finite time. We then proceed to consider travelling wave solutions of the same system and establish the existence of such a solution.

Key words. free boundary, reaction-diffusion equations, travelling wave solutions, wound healing

AMS subject classifications. 35K57, 35R35, 92C50

PII. S0036141099351693

Introduction. The mathematical modeling of wound healing has received increased attention in recent years. One area of investigation is dermal wound healing, where complex biological processes are interacting [4, 7]. Two of these processes are the invasion of fibroblasts into the wound space and their alignment there [5], and the sprouting of blood vessels into the wound space (angiogenesis) [6]. Another area of investigation is corneal epithelial wound healing, which was modelled by Dale, Maini, and Sheratt [1]. The modeling framework involves two concentrations: the generic corneal epithelial cell density N (also referred to as the corneal stimulus) in the healed region, and a chemical stimulus concentration C (also called the epidermal growth factor) both in the healed and the wound regions. The epithelial cells are assumed to migrate as a diffusing substance. The chemical stimulus also diffuses, by mechanisms such as tear fluid convection and mixing arising from blinking. Thus the model involves a system of two parabolic partial differential equations (semilinear reaction-diffusion equations) for the concentration N, in the healed region, and the concentration C, both in the healed and the wound regions. The model developed in [1] is one-dimensional and it exhibits, after a short time, a travelling wave character with speed approximately 20 μmh^{-1}; the actual observed healing rate in corneal wounds is approximately 60 μmh^{-1}.

This model was recently improved by Gaffney et al. [2] in two ways. First, they include the presence of a physiological electric field and, second, they introduce a free boundary into the model. The free boundary is the receding boundary of the wound region. The physiological electric field arises from transcornea potential difference near the boundary of the healed region: whereas the cells in the healed region maintain normal potential, in the wound region the potential is short-circuited. The physiological electric field is most significant near the free boundary. It increases the transport of epithelial cells into the wound, thus increasing the speed of the healing

*Received by the editors February 5, 1999; accepted for publication (in revised form) July 25, 2000; published electronically December 5, 2000.
http://www.siam.org/journals/sima/32-4/35169.html
†Department of Mathematics, University of Pittsburgh, Pittsburgh, PA 15260 (xinfu@pitt.edu). The first author was partially supported by National Science Foundation grant DMS-962287.
‡Department of Mathematics, University of Minnesota, 206 Church Street SE, Minneapolis, MN 55455 (friedman@ima.umn.edu). The second author was partially supported by National Science Foundation grant DMS 94-01251.

process. The computations in [2] show that the average speed of the free boundary varies linearly with the electric field over a large range, and that this linear relation is robust under variations of some parameters which are difficult to estimate precisely.

The purpose of this paper is to study the model introduced in [2] by rigorous mathematical analysis. Our study consists of two parts, which are, technically, quite different. In the first part (Part I) we study the evolution of the free boundary problem for the system of the two reaction-diffusion equations for N and C and prove that the problem is well posed, and that complete healing is achieved in finite time. In the second part (Part II) we consider the speed of the healing process. As mentioned above, numerical computations in [1, 2] show that healing proceeds (after a short initial time) with constant speed, as a travelling wave. We prove, by rigorous analysis, that a travelling wave solution indeed exists.

In order to make the paper more readable, we shall deal with a general system of reaction-diffusion equations, for concentrations c and n, subject to some general assumptions. In the concluding section of the paper we write down the explicit system developed in [2], and show that it satisfies the general assumptions made throughout Parts I and II.

Part I: The evolution problem.

1. The model. Consider the following system for $n(x,t)$, $c(x,t)$, $s(t)$:

$$n_t = (d(n,c)n_x)_x + f(n,c), \quad 0 < x < s(t),\ t > 0, \tag{1.1}$$

$$ac_t = c_{xx} + g(n,c), \quad 0 < x < 1,\ t > 0 \quad (a \text{ a positive constant}), \tag{1.2}$$

$$n \equiv 0 \quad \text{if } s(t) < x < 1,\ t > 0, \tag{1.3}$$

$$n(s(t),t) = n_*, \quad t > 0 \quad (n_* \text{ constant}), \tag{1.4}$$

$$\dot{s}(t) = -\frac{1}{n_*} d(n_*,c)n_x \quad \text{at } x = s(t),\ t > 0, \tag{1.5}$$

with boundary conditions

$$n_x(0,t) = 0, \quad c_x(0,t) = 0, \quad c_x(1,t) = 0 \quad \text{for } t > 0 \tag{1.6}$$

and initial conditions

$$\begin{aligned} s(0) &= s_0, & 0 < s_0 < 1, \\ n(x,0) &= n_0(x), & 0 < x < s_0, \\ c(x,t) &= c_0(x), & 0 < x < 1. \end{aligned} \tag{1.7}$$

Here n represents the corneal stimulus cell density in the healed region $0 \le x < s(t)$, and c represents the chemical stimulus concentration in both the wound and the healed regions.

In the condition (1.4), n_* is the equilibrium cell density at the boundary of the wound. The condition (1.5) is a conservation law of cell mass; it says that the rate of increase in the cell mass at the free boundary, namely, $n_*\dot{s}(t)$, is equal to the flux of cells, $d(n_*,c)n_x$. This condition is the same as the classical Stefan condition in the model of melting of solids. As will be shown in section 9, the term $d(n,c)n_x$ in (1.1) includes the effect of diffusion due to the physiological electric field mentioned in the introduction, and the functions $f(n,c)$, $g(n,c)$ account for chemical effects and sources for c and n.

We assume that

$$0 < n_* < 1 \tag{1.8}$$

and set

$$G = \{(n, c); \quad n^* \le n \le 2, \quad 0 \le c \le 1\}.$$

We also make the following assumptions:

$$\begin{aligned} &d(n,c) \text{ is in } C^1(G) \quad \text{and} \quad \partial d/\partial n \le 0, \\ &d(n,c) \ge d_0 > 0 \quad (d_0 \text{ constant}); \end{aligned} \tag{1.9}$$

$$\begin{aligned} &f(n,c) \text{ is in } C^1(G) \quad \text{and} \\ &f(n_*, c) > 0, \quad f(2,c) \le 0 \quad \text{if } 0 \le c \le 1; \end{aligned} \tag{1.10}$$

$$\begin{aligned} &g(n,c) \text{ is in } C^1(G) \quad \text{and} \quad \partial g/\partial c < 0 \text{ in } G, \\ &g(n,0) \ge 0, \quad g(n,1) \le 0 \quad \text{for } n_* \le n \le 2; \end{aligned} \tag{1.11}$$

$$\begin{aligned} &n_0 \in C^2[0, s_0], \quad n_* < n_0(x) < 2 \quad \text{if } 0 \le x < s_0, \\ &n_0'(0) = 0, \quad n_0(s_0) = n_*, \quad n_0'(s_0) < 0; \end{aligned} \tag{1.12}$$

$$c_0 \in C^2[0,1] \quad \text{and} \quad 0 \le c_0(x) \le 1. \tag{1.13}$$

We may clearly extend the definitions of the functions d, f, g in such a way that

$$\begin{aligned} &(1.9), (1.10), (1.11) \text{ hold for } -\infty < n, \ c < \infty, \text{ and} \\ &f(n,c) > 0 \quad \text{if } n \le n_*, \quad f(n,c) \le 0 \quad \text{if } n \ge 2 \quad \text{for all } c, \\ &g(n,c) \ge 0 \quad \text{if } c \le 0, \quad g(n,c) \le 0 \quad \text{if } c \ge 1 \quad \text{for all } n. \end{aligned} \tag{1.14}$$

In the future, we shall assume such an extension has already been made.

In sections 2 and 3, we shall prove the following theorem.

THEOREM 1.1. *Assume that* (1.8)–(1.13) *hold. Then there exists a unique solution* (n, c, s) *of* (1.1)–(1.7) *for* $0 \le t < T_*$, *where* $T_* < \infty$, *and*

$$s'(t) > 0 \quad \textit{if } 0 \le t < T_*, \tag{1.15}$$

$$s(t) \to 1 \quad \textit{if } t \to T_*; \tag{1.16}$$

furthermore,

$$n_* < n(x,t) \le 2 \quad \textit{if } 0 \le x < s(t), \quad 0 \le t < T_*, \tag{1.17}$$

$$0 \le c(x,t) \le 1 \quad \textit{if } 0 \le x \le 1, \qquad 0 \le t < T_*. \tag{1.18}$$

Note that (1.16) means that the wound region $\{s(t) \le x \le 1\}$ disappears as $t \to T_*$.

In section 2, we shall prove local existence and uniqueness. In section 3, we shall derive a priori estimates that will enable us to complete the proof of Theorem 1.1.

2. Local existence and uniqueness. In this section, we prove the following lemma.

LEMMA 2.1. *Under the assumptions of Theorem* 1.1, *there exists a unique solution of* (1.1)–(1.7) *for a small time interval* $0 \le t \le T$ $(T > 0)$.

Proof. Set

$$\delta = \min\{s_0, 1 - s_0\}$$

and let $\zeta(y)$ be a function in $C^3[0,1]$ satisfying

$$\zeta(y) = 1 \quad \text{if } |y - s_0| < \frac{\delta}{4}, \qquad \zeta(y) = 0 \quad \text{if } |y - s_0| > \delta, \qquad |\zeta'(y)| < \frac{2}{\delta}.$$

We introduce a transformation that will straighten the free boundary:

$$(x,t) \to (y,t), \quad \text{where } x = y + \zeta(y)(s(t) - s_0), \quad 0 \le y \le 1. \tag{2.1}$$

Notice that as long as

$$|s(t) - s_0| < \frac{\delta}{4} \tag{2.2}$$

the transformation (2.1) is a diffeomorphism from $[0,1]$ onto $[0,1]$ (since $\partial x/\partial y > \frac{1}{2}$), and

$$\begin{aligned} 0 \le x \le s(t) &\iff 0 \le y \le s_0, \\ s(t) \le x \le 1 &\iff s_0 \le y \le 1, \\ x = s(t) &\iff y = s_0. \end{aligned}$$

One easily computes that

$$\begin{aligned} \frac{\partial y}{\partial x} &= \frac{1}{1 + \zeta'(y)(s(t) - s_0)} \equiv \sqrt{A(s(t), y)}, \\ \frac{\partial^2 y}{\partial x^2} &= \frac{\zeta''(y)(s(t) - s_0)}{[1 + \zeta'(y)(s(t) - s_0)]^3} \equiv B(s(t), y), \\ -\frac{1}{\dot{s}(t)} \frac{\partial y}{\partial t} &= \frac{\zeta(y)}{1 + \zeta'(y)(s(t) - s_0)} \equiv C(s(t), y). \end{aligned}$$

Defining

$$\varphi(y,t) = n(x,t), \qquad \psi(y,t) = c(x,t)$$

and setting

$$D(\varphi,\psi) = d(n,c), \qquad F(\varphi,\psi) = f(n,c), \qquad G(\varphi,\psi) = g(n,c),$$

the system (1.1)–(1.7) takes the form

$$\varphi_t = AD(\varphi,\psi)\varphi_{yy} + (BD(\varphi,\psi) + \dot{s}C)\varphi_y + AD_\varphi \varphi_y^2 + AD_\psi \varphi_y \psi_y + F(\varphi,\psi) \quad \text{for } 0 < y < s_0,\ t > 0, \tag{2.3}$$

$$a\psi_t = A\psi_{yy} + (B + a\dot{s}C)\psi_y + G(\varphi,\psi) \quad \text{for } 0 < y < 1,\ t > 0, \tag{2.4}$$

$$\varphi \equiv 0 \quad \text{if } s_0 < y < 1, \tag{2.5}$$

with $A = A(s,y)$, $B = B(s,y)$, $C = C(s,y)$,

$$\varphi(s_0,t) = n_*, \tag{2.6}$$

$$\dot{s}(t) = -\frac{1}{n_*} D(n_*, \psi(s_0,t))\varphi_y(s_0,t), \tag{2.7}$$

and

$$\varphi_y(0,t) = \psi_y(0,t) = \psi_y(1,t) = 0, \tag{2.8}$$

$$\begin{cases} \varphi(y,0) = \varphi_0(y), & 0 \le y < s_0, \\ \psi(y,0) = \psi_0(y), & 0 \le y \le 1, \end{cases} \tag{2.9}$$

where $\varphi_0(y) = n_0(x)$, $\psi_0(y) = c_0(x)$.

We introduce the quantity

$$s_1 = -\frac{1}{n_*} D(n_*, \psi_0(s_0))\varphi_0'(s_0) \quad (s_1 > 0), \tag{2.10}$$

which should be the derivative $\dot{s}(0)$ if a solution exists.

We shall prove existence by invoking the Schauder fixed point theorem. Toward this purpose, we introduce spaces

$$X_T = \{s \in C^1[0,T], \quad s(0) = 0, \quad \dot{s}(0) = s_1, \quad |\dot{s}(t) - s_1| \le 1 \quad \text{for } 0 \le t \le T\},$$
$$Y_T = \{\varphi \in C^0([0,s_0] \times [0,T]), \quad \varphi(y,0) = \varphi_0, \quad |\varphi - \varphi_0|_{C^0([0,s_0]\times[0,T])} \le 1\},$$

where T is such that

$$0 < T < \frac{\delta}{4(1+s_1)}.$$

For any $(\tilde{s}, \tilde{\varphi}) \in X_T \times Y_T$ we then have

$$|\tilde{s}(t) - s_0| < \frac{\delta}{4},$$

so that the mapping $(x,y) \to (y,t)$ defined by

$$x = y + \zeta(y)(\tilde{s}(t) - s_0)$$

is a diffeomorphism, and we define ψ to be the solution of

$$a\psi_t = A(\tilde{s},y)\psi_{yy} + (B(\tilde{s},y) + a\dot{\tilde{s}}C(\tilde{s},y))\psi_y + G(\tilde{\varphi},\psi) \quad \text{for } 0 < y < 1,\ 0 < t < T$$

with $\tilde{\varphi} \equiv 0$ in $[s_0,1] \times [0,T]$,

$$\psi_x(0,t) = \psi_x(1,t) = 0, \quad 0 < t < T,$$
$$\psi(y,0) = \psi_0(y).$$

Using L^p estimates for parabolic equations and Sobolev's inequalities, one can show that this system has a unique solution ψ with finite norm

$$||\psi||_{C^{1+\beta,(1+\beta)/2}([0,1]\times[0,T])} \le K \tag{2.11}$$

for any $0 < \beta < 1$, where K will be used to denote constants depending only on s_0, s_1, and on

$$|\psi_0|_{C^2[0,s_0]}, \quad |\psi_0|_{C^2[0,1]}.$$

In fact, results of this type, for more general nonlinear parabolic equations, are proved in [3].

Next, we define a function φ as the solution of the parabolic problem

$$\begin{aligned}
\varphi_t &= A(\tilde{s},y)D(\tilde{\varphi},\psi)\varphi_{yy} + [B(\tilde{s},y)D(\tilde{\varphi},\psi) + \dot{\tilde{s}}C(\tilde{s},y)]\varphi_y + A(\tilde{s},y)D_\varphi(\tilde{\varphi},\psi)\varphi_y^2 \\
&\quad + A(\tilde{s},y)D_\psi(\tilde{\varphi},\psi)\varphi_y\psi_y + f(\varphi,\psi) \quad \text{for } 0 \le y \le s_0,\ 0 \le t \le T, \\
\varphi_y(0,t) &= 0, \quad \varphi(s_0,t) = n_* \quad \text{for } 0 < t < T, \\
\varphi(y,0) &= \varphi_0(y) \quad \text{for } 0 < y < s_0.
\end{aligned}$$

As before, this system has a unique solution φ with finite norm

$$||\varphi||_{C^{1+\beta,(1+\beta)/2}([0,s_0]\times[0,T])} \le K. \tag{2.12}$$

Finally, we define

$$s(t) = s_0 - \int_0^t \frac{1}{n_*} D(n_*,\psi(s_0,\tau))\varphi_y(s_0,\tau)d\tau \tag{2.13}$$

and introduce the mapping W by

$$(s,\varphi) = W(\tilde{s},\tilde{\varphi}).$$

We want to show that W has a fixed point (s,φ), which will then imply that, together with the corresponding ψ, it forms a solution to the system (2.3)–(2.9).

Observe that

$$\dot{s}(t) - s_1 = \frac{1}{n_*}\left\{D(n_*,\psi_0(s_0))\frac{\partial\varphi_0(s_0)}{\partial y} - D(n_*,\psi(s_0,t))\varphi_y(s_0,t)\right\}$$

and the right-hand side is in $C^{\beta/2}[0,T]$. Hence

$$||\dot{s} - s_1||_{C^{\beta/2}[0,T]} \le K. \tag{2.14}$$

By (2.12), we also have

$$||\varphi - \varphi_0||_{C^{1+\beta,(1+\beta)/2}([0,s_0]\times[0,T])} \le K.$$

Hence, if T is small enough,

$$\begin{aligned}
||\dot{s} - s_1||_{C^0[0,T]} &\le ||\dot{s} - s_1||_{C^{\beta/2}[0,T]} T^{\beta/2} \le 1, \\
||\varphi - \varphi_0||_{C^0[0,s_0]} &\le ||\varphi - \varphi_0||_{C^{0,(1+\beta)/2}([0,s_0]\times[0,T])} T^{\frac{1+\beta}{2}} \le 1,
\end{aligned}$$

so that W maps $X_T \times Y_T$ into itself.

From the estimates above, it follows that the image of W lies in a compact subset of $X_T \times Y_T$, and a standard argument then also shows that W is continuous. Invoking the Schauder fixed point theorem we conclude that W has a fixed point in $X_T \times Y_T$.

We can further use the Schauder estimates to obtain additional regularity of the solution, such as the Hölder continuity of $\dot{s}(t)$, and of the second spatial derivatives of φ and ψ. It remains to prove uniqueness.

Let (s_i, φ_i, ψ_i) $(i = 1, 2)$ be two solutions. Then

$$||s_i||_{C^{1+\beta/2}[0,T]} + ||\varphi_i||_{C^{1+\beta,(1+\beta)/2}([0,s_0]\times[0,T])} + ||\psi_i||_{C^{1+\beta,(1+\beta)/2}([0,1]\times[0,T])} \le K.$$

Setting $\psi = \psi_1 - \psi_2$ and taking the difference of the equations for ψ_1, ψ_2, we get

$$\begin{aligned} a\psi_t = {} & A(s_2, y)\psi_{yy} + [B(s_2, y) + a\dot{s}_2 C(s_2, y)]\psi_y + G_1\varphi + G_2\psi \\ & + [A(s_1, y) - A(s_2, y)]\psi_{1,yy} + [B(s_1, y) - B(s_2, y)]\psi_{1,y} \\ & + a[\dot{s}_1 C(s_1, y) - \dot{s}_2 C(s_2, y)]\psi_{1,y}, \end{aligned}$$

where G_1, G_2 are the partial derivatives of G with respect to the first and second variables evaluated at an intermediate point.

Using $W^{2,p}$ estimates for parabolic equations of the form $u_t - au_{yy} - bu_y + cu = g$ and Sobolev's imbedding, we get

$$||\psi_1 - \psi_2||_{C^{1+\beta,(1+\beta)/2}} \le K\{||\varphi_1 - \varphi_2||_{C^0} + ||s_1 - s_2||_{C^1}\}. \tag{2.15}$$

Similarly, we derive the inequality

$$||\varphi_1 - \varphi_2||_{C^{1+\beta,(1+\beta)/2}} \le K\{||\psi_1 - \psi_2||_{C^{1,0}} + ||s_1 - s_2||_{C^1}\}. \tag{2.16}$$

Furthermore, taking the difference of the equations for s_1, s_2 in (2.7), we get

$$||s_1 - s_2||_{C^{1+\beta/2}} \le K||\psi_1 - \psi_2||_{C^{0,\beta/2}} + ||\varphi_{1,y} - \varphi_{2,y}||_{C^{0,\beta/2}}. \tag{2.17}$$

It then follows, upon using (2.16), (2.15), that

$$\begin{aligned} & ||s_1 - s_2||_{C^{1+\beta/2}} + ||\varphi_1 - \varphi_2||_{C^{1+\beta,(1+\beta)/2}} + ||\psi_1 - \psi_2||_{C^{1+\beta,(1+\beta)/2}} \\ & \quad \le K\{||\varphi_1 - \varphi_2||_{C^0} + ||s_1 - s_2||_{C^1}\} \\ & \quad \le KT^{\beta/2}\{||\varphi_1 - \varphi_2||_{C^{1+\beta,(1+\beta)/2}} + ||s_1 - s_2||_{C^{1+\beta/2}}\}. \end{aligned}$$

Taking T such that also $KT^{\beta/2} < 1$, we conclude that $s_1 \equiv s_2$, $\varphi_1 \equiv \varphi_2$, $\psi_1 \equiv \psi_2$. □

3. Completion of the proof of Theorem 1.1. We first derive a priori estimates for any solution of (1.1)–(1.7), assuming that it exists in some interval $0 < t < T$; these bounds will be independent of T.

LEMMA 3.1. *The solution satisfies:*

$$n_* < n(x, t) < 2 \quad \textit{if } 0 \le x < s(t),\ 0 \le t \le T, \tag{3.1}$$

$$0 \le c(x, t) \le 1 \quad \textit{if } 0 \le x \le 1,\ 0 \le t \le T, \tag{3.2}$$

and

$$n_x(s(t), t) < 0 \quad \textit{if } 0 \le t \le T. \tag{3.3}$$

Proof. By (1.14),

$$n_t - (dn_x)_x \ge 0 \quad \text{if } n \le n_*.$$

Hence the maximum principle yields the inequality $n > n_*$ if $0 < x < s(t)$, $0 < t < T$. Similarly, we have from (1.14) that

$$n_t - (dn_x)_x \le 0 \quad \text{if } n \ge 2$$

so that, by the maximum principle, $n < 2$ if $0 \le x \le s(t)$, $0 \le t \le T$. The proof of (3.2) is similar. Finally, noting that, since $f(n_*, c) > 0$,

$$n_t - (dn_x)_x = f(n, c) > 0$$

near the free boundary, the (strict) inequality (3.3) follows by the maximum principle. □

From (3.3) and (1.5), we deduce that

$$\dot{s}(t) > 0 \quad \text{for } 0 \le t \le T \tag{3.4}$$

so that $s(t)$ is strictly monotone increasing.

Also, a standard parabolic estimate shows that c_x is bounded independently of T.

LEMMA 3.2. *There exists a constant M independent of T such that*

$$\dot{s}(t) \le M \quad \textit{for } 0 < t < T. \tag{3.5}$$

Proof. Let

$$Q = \{(x, t); \quad 0 < x < s(t), \quad 0 < t < T\}$$

and introduce the operator

$$\mathcal{L}u = u_t - d(u, c)u_{xx} - d_u(u, c)u_x^2 - d_c(u, c)u_x c_x - f(u, c),$$

where c is the component of the solution (s, n, c) of (1.1)–(1.7). We shall construct a supersolution in the form

$$\bar{n}(x, t) = \begin{cases} 2, & 0 < x < s(t) - \dfrac{1}{M}, \\ n_* + (2 - n_*)\{2M(s(t) - x) - M^2(s(t) - x)^2\}, & s(t) - \dfrac{1}{M} < x < s(t). \end{cases}$$

Notice that $\bar{n} \in W_\infty^{2,1}(Q)$ and

$$\bar{n}_x \le 0, \qquad \bar{n}(s(t), t) = n_*.$$

In the interval $[0, s(t) - \frac{1}{M})$

$$\mathcal{L}\bar{n} = -f(2, c) \ge 0.$$

In the interval $(s(t) - \frac{1}{M}, s(t))$

$$\begin{aligned} \bar{n}_t &= (2 - n_*)2M\dot{s}\{1 - M(s - x)\} \ge 0 \quad \text{since } \dot{s}(t) > 0, \\ -d(\bar{n}, c)\bar{n}_{xx} &= (2 - n_*)2M^2 d(\bar{n}, c) \ge c_0 M^2, \end{aligned}$$

where $c_0 = (2 - n_*)2d_0 > 0$ (see (1.9)),

$$\begin{aligned} -d_n(\bar{n}, c)\bar{n}_x^2 &\ge 0 \qquad \text{(by (1.9))}, \\ |-d_c(\bar{n}, c)c_x\bar{n}_x| &\le c_1 M, \end{aligned}$$

and

$$|f(\bar{n}, c)| \le c_2,$$

where c_1, c_2 are constants independent of M and T. It follows that

$$\mathcal{L}\bar{n} \ge c_0 M^2 - c_1 M - c_2 > 0$$

if M is large enough. Furthermore,

$$\bar{n}_x = 0 = n_x \quad \text{at } x = 0, \qquad \bar{n} = n_* = n \quad \text{at } x = s(t),$$

and $\bar{n}(x, 0) \ge n_0(x)$ for large M. Hence by comparison, $n(x, t) \le \bar{n}(x, t)$ in Q and, consequently,

$$n_x(s(t), t) \ge \bar{n}_x(s(t), t) = -2M(2 - n_*).$$

The assertion (3.5) (with another M) now follows by recalling (1.5).

From (3.5) and (3.4), we get

$$|\dot{s}(t)| \le M. \qquad \square \tag{3.6}$$

LEMMA 3.3. *The solution to* (1.1)–(1.7) *exists and is unique, and it can be extended up to a time* T^* *satisfying* $\lim_{t \nearrow T^*} s(t) = 1$.

Proof. The assertion means that if the solution exists for $0 < t < T_0$, and if $s(T_0) < 1$, then the solution can be continued, uniquely, to a larger interval $0 < t < T_0 + \tau$ ($\tau > 0$). To prove this we observe that in the proof of Lemma 2.1 the size T of the time interval depends on a lower bound on $\min\{s, 1 - s\}$ at the initial time and on the L^∞ bound of the first two derivatives of the initial data. From (3.6) and L^p estimates applied to the system with "straightened" free boundary, (2.3)–(2.9), we deduce a priori bounds on the first two derivatives of $n(x, t)$, $c(x, t)$ at $t = T_0$ (in fact, even $C^{2+\alpha}$ bounds if we use a bootstrap argument and Schauder's estimate). Since, further, $0 < 1 - s(T_0) < 1$, we can extend the solution, and uniquely so, to a larger interval $0 < t < T_0 + \tau$, as claimed. $\square$

We next improve inequality (3.4).

LEMMA 3.4. *There exists a positive constant* γ *independent of* T *such that*

$$\dot{s}(t) \ge \gamma \quad \textit{for } 0 \le t < T. \tag{3.7}$$

Proof. We shall construct a subsolution in the form

$$\underline{n}(x, t) = n_* + \begin{cases} \dfrac{\gamma}{2} s_0^2, & 0 < x < s(t) - s_0, \\ \gamma \left[s_0(s(t) - x) - \dfrac{1}{2}(x - s(t))^2 \right], & s(t) - s_0 < x < s(t). \end{cases}$$

Notice that $\underline{n} \in W^{2,1}_\infty(G)$.

In the interval $(0, s(t) - s_0)$,

$$\mathcal{L}\underline{n} = -f\left(n_* + \frac{\gamma}{2} s_0^2, c\right) < 0$$

by (1.10) if γ is sufficiently small. In the interval $(s(t) - s_0, s(t))$,

$$\underline{n}_t = \gamma[s_0 - (s(t) - x)]\dot{s}(t) \le \gamma M \qquad \text{(by (3.6))},$$
$$f(\underline{n}, c) \ge f(n_*, c) - C_1 \gamma,$$

and all the other terms in $\mathcal{L}\underline{n}$ are bounded by $C_2\gamma$, where C_1, C_2 are positive constants independent of γ and T. Since by (1.10), $f(n_*, c) \geq \mu > 0$ for all c, we conclude that

$$\mathcal{L}\underline{n} \leq -\mu + \gamma(M + C_1 + C_2) < 0$$

if γ is sufficiently small. We also have

$$\underline{n}_x = 0 = n_* \quad \text{at } x = 0, \qquad \underline{n} = n_* = n \quad \text{at } x = s(t)$$

and $\underline{n}(x, 0) < n_0(x)$ if $0 \leq x < s_0$ provided γ is small enough; here we used the assumption $n_0'(s_0) < 0$.

By comparison we then have $\underline{n} \leq n$ in G and

$$n_x(s(t), t) \leq \underline{n}_x(s(t), t) = -\gamma s_0$$

so that (3.7) holds (with another γ).

Combining Lemma 3.3 with Lemma 3.4, we see that there exists a finite number T_* such that the solution exists for all $0 < t < T_*$, and $s(T_*) = 1$. This completes the proof of Theorem 1.1. □

Part II: Travelling wave solutions.

4. Setting up the problem. We seek a solution to (1.1)–(1.5) in the form of a travelling wave with constant speed c ($c > 0$):

$$n(x, t) = u(z), \qquad c(x, t) = v(z), \qquad z = ct - x \in \mathbb{R}^1.$$

The free boundary is given by $x = ct$, i.e., $z = 0$. Then

$$(d(u, v)u')' - cu' + f(u, v) = 0, \quad z > 0, \tag{4.1}$$

$$v'' - acv' + g(u, v) = 0, \quad -\infty < z < \infty, \tag{4.2}$$

$$u(z) \equiv 0, \quad z < 0. \tag{4.3}$$

Motivated by numerical results from [2], we wish to consider only solutions such that

$$u'(z) > 0, \quad z > 0. \tag{4.4}$$

We impose the boundary conditions

$$u(0) = u_* \quad (u_* = n_*), \tag{4.5}$$

$$d(u(0), v(0))u'(0) = cu_*, \tag{4.6}$$

$$u(\infty) = 1, \tag{4.7}$$

and

$$v(-\infty) = 1, \tag{4.8}$$

$$v(+\infty) = 0. \tag{4.9}$$

As a first step, we shall simplify the problem by reducing it to a system in the interval $\{z > 0\}$ only.

Set

$$g_0(v) = g(0, v), \qquad \beta = ac \tag{4.10}$$

so that

$$v'' - \beta v' + g_0(v) = 0 \quad \text{for } -\infty < z < 0. \tag{4.11}$$

LEMMA 4.1. *Assume that $g_0(v)$ satisfies*

$$g_0(1) = 0, \qquad \frac{d}{dv} g_0(v) < 0 \quad \textit{for all } v \in \mathbb{R}^1.$$

Then there exists a smooth function $\Psi(\beta, v)$ such that for every $\beta \in (0, \infty)$, the problem

$$v'' - \beta v' + g_0(v) = 0 \quad \textit{in } (-\infty, 0), \qquad v(0) = v_0, \qquad v'(0) = v_1 \tag{4.12}$$

has a bounded solution if and only if

$$v_1 = \Psi(\beta, v_0); \tag{4.13}$$

furthermore, the bounded solution satisfies $\lim_{z \to -\infty} v(z) = 1$. The function $\Psi(\beta, v)$ has the following properties:

$$\begin{aligned} \frac{d\Psi}{dv} &= \beta - \tfrac{g_0(v)}{\Psi} \quad \textit{in } \mathbb{R}^1 \backslash \{1\}, \\ \Psi(\beta, 1) &= 0, \end{aligned} \tag{4.14}$$

and

$$\Psi_v > 0 \quad \textit{in } \mathbb{R}^1. \tag{4.15}$$

Proof. We write (4.11) as an autonomous system

$$v' = p, \qquad p' = \beta p - g_0(v).$$

In the $v - p$ phase plane there is only one stationary point, namely, (1,0). It is a saddle point since the characteristic equation

$$\lambda^2 - \beta\lambda + g_0'(1) = 0$$

has two real roots, one positive and one negative. Hence there are only two trajectories leaving (1,0), which we shall denote by γ^+ and γ^-. The velocity field of the autonomous system is shown in Figure 1. Note that every trajectory can intersect the v-axis at most once. Hence $p = v'$ can change sign at most once and, consequently, $v(-\infty)$ exists. Since there is only one stationary point, we conclude that either $|v(-\infty)| = \infty$ or $v(-\infty) = 0$. Hence a solution is bounded if and only if (v_0, v_1) lies on the curve $\gamma^+ \cup \{(1,0)\} \cup \gamma^- \equiv \Gamma$.

Along γ^+, $p > 0$ and $v > 1$ so that $p' = \beta p - g_0(v) > 0$, whereas along γ^-, $p < 0$ and $v < 1$ so that $p' < 0$. It follows that Γ can be written as a curve $p = \Psi(\beta, v)$ and, then, a solution is bounded if and only if $v_1 = \Psi(\beta, v_0)$. It is clear that $\Psi(\beta, 1) = 0$, and the equation in (4.14) follows from

$$\frac{dp}{dv} = \beta - \frac{g_0}{p}.$$

From the last equation we get

$$\frac{d}{dv}\left(p \frac{dp}{dv}\right) = \beta \frac{dp}{dv} - \frac{dg_0}{dv} > \beta \frac{dp}{dv}. \tag{4.16}$$

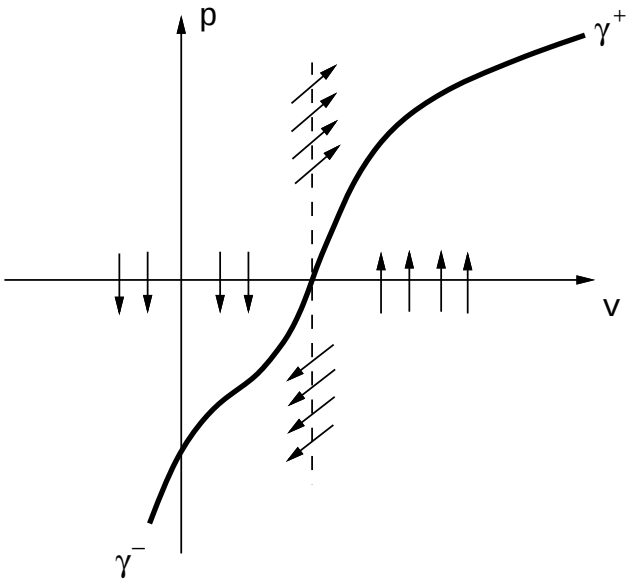


FIG. 1.

To prove (4.15) notice that the inequality holds in a neighborhood of the origin. If the inequality does not hold for all $v \neq 0$ then there is a smallest positive v^* (or a largest negative value v^{**}) at which $\Psi_v(\beta, v)$ vanishes. It follows that $\Psi_{vv} \leq 0$ at v^* (or $\Psi_{vv} \geq 0$ at v^{**}). We also have, however, $p = p(v^*) > 0$ ($p = p(v^{**}) < 0$); cf. Figure 1. Therefore, the left-hand side of (4.16) (with $p = \Psi(\beta, v)$) is negative, whereas the right-hand side vanishes at v^* (or v^{**}), which is a contradiction. □

Remark 4.1. In the special case $g_0(v) = k(1 - v)$ for some $k > 0$,

$$v = 1 - Ce^{\lambda z},$$

where

$$\lambda = \frac{1}{2}(\beta + \sqrt{\beta^2 + 4k}),$$

and $\Psi(\beta, v) = \lambda(v - 1)$.

Lemma 4.1 allows us to reformulate problem (4.1)–(4.9) as follows.

Problem (P_0). Find (c, u, v) with $c > 0$ such that

$$(d(u,v)u')' - cu' + f(u,v) = 0, \qquad z > 0, \tag{4.17}$$

$$v'' - acv' + g(u,v) = 0, \qquad z > 0, \tag{4.18}$$

$$u(0) = u_*, \quad d(u_*, v(0))u'(0) = cu_*, \quad u(\infty) = 1, \tag{4.19}$$

$$-v'(0) + \Psi(ac, v(0)) = 0, \qquad v(\infty) = 0, \tag{4.20}$$

$$u'(z) > 0, \qquad z > 0. \tag{4.21}$$

From (4.21) and (4.19), we see that as z varies from 0 to ∞, $u = u(z)$ varies from u_* to 1. We can therefore try to use u as an independent variable, i.e., set $z = z(u)$ as the inverse function of $u = u(z)$; u will vary in the interval

$$\Omega = \{u_* < u < 1\}.$$

We introduce new dependent functions by

$$Q(u) = d(u,v)u'(z)|_{z=z(u)}, \qquad V(u) = v(z)|_{z=z(u)}. \tag{4.22}$$

Using the rule

$$\frac{d}{dz} = u'\frac{d}{du} = \frac{Q(u)}{d(u,v)}\frac{d}{du},$$

we can transform problem (P_0) into the following problem.

Problem (P_1). Find (c, Q, V) with $c > 0$ such that

$$QQ_u - cQ + d(u,V)f(u,V) = 0 \quad \text{in } \Omega, \tag{4.23}$$

$$\frac{Q}{d(u,V)}\left(\frac{Q}{d(u,V)}V_u\right)_u - ac\frac{Q}{d(u,V)}V_u + g(u,V) = 0 \quad \text{in } \Omega, \tag{4.24}$$

$$Q(u_*) = cu_*, \quad Q(1) = 0, \tag{4.25}$$

$$-\frac{Q(u_*)}{d(u_*, V(u_*))}V_u + \Psi(ac, V(u_*)) = 0, \quad V(1) = 0, \tag{4.26}$$

$$Q > 0 \quad \text{in } \Omega, \tag{4.27}$$

$$\int_{u_*}^{1} \frac{d(s,V(s))}{Q(s)}\,ds = \infty. \tag{4.28}$$

Indeed, we have the following lemma.

LEMMA 4.2. *Problems* (P_0) *and* (P_1) *are equivalent.*

Proof. Given a solution to problem (P_0), define $z = z(u)$ as the inverse of $u = u(z)$ and define Q, V by (4.22). Then (4.23), (4.24), (4.26), (4.27), and the first boundary condition in (4.25) hold. Since

$$z = \int_{u_*}^{u} \frac{d(s,V(s))}{Q(s)}\,ds \tag{4.29}$$

and $u(\infty) = 1$, we also have (4.28). Finally, to prove that $Q(1) = 0$, note that the existence of $u(\infty)$ and $v(\infty)$ implies by standard ODE theory that u'', v'' are bounded, and then, by a simple argument in real analysis, $u'(\infty), v'(\infty)$ must exist and be equal to zero. Taking $u \to 1$ in the first relation of (4.22), we then deduce that $Q(1) = 0$.

Conversely, consider a solution to problem (P_1) and define $u = u(z)$ by

$$u'(z) = \frac{Q(u)}{d(u,V(u))}, \quad u(0) = u_* \tag{4.30}$$

and $v(z) = V(u(z))$. Then (4.29) holds, and one can easily show that all the equations in problem (P_0) are satisfied. ∎

5. Statement of the main result. In what follows, we shall work primarily (but not exclusively) with problem (P_1). We shall need some assumptions that include all those made in (1.9)–(1.11) (except for the condition $f(2,c) \le 0$, which will not be needed). Using the variables (u,v) instead of (n,c), we introduce the set (which coincides with G)

$$S = \{(u,v); \quad u_* \le u \le 1, \quad 0 \le v \le 1\}.$$

We assume that

$$\begin{aligned}&\text{the function } d(u,v) \text{ is in } C^1(S), \quad \text{and}\\ &0 < d_0 \le d(u,v) \le d_1 < \infty, \quad d_u \le 0, \quad d_v \ge 0, \quad ad \le 1,\end{aligned} \tag{5.1}$$

where a is the constant appearing in (4.2) and (4.18);

$$\begin{aligned}&\text{the function } f(u,v) \text{ is in } C^1(S) \quad \text{and}\\ &f(1,0) = 0, \quad f_u(1,0) < 0, \quad f(u,0) > 0 \quad \text{for } u_* \le u < 1,\\ &f_v(u,v) > 0 \quad \text{in } S;\end{aligned} \tag{5.2}$$

$$\begin{aligned}&\text{the function } g(u,v) \text{ is in } C^1(S) \quad \text{and}\\ &g(1,0) = 0, \quad g(u,0) \ge 0, \quad \text{and} \quad g(u,1) \le 0 \quad \text{for } u_* \le u \le 1,\\ &g_u(u,v) < 0, \quad g_v(u,v) < 0 \quad \text{in } S;\end{aligned} \tag{5.3}$$

$$\begin{aligned}&\text{there exists a positive number } \ell \text{ such that } \ell(1-u_*) > 1 \quad \text{and}\\ &\max_{0\le v\le \ell(1-u)} \frac{f(u,v)}{d(u,v)} + \frac{1}{\ell} g(u, \ell(1-u)) \le 0 \quad \text{for } 1 - \frac{1}{\ell} \le u \le 1.\end{aligned} \tag{5.4}$$

Note that $1 - \frac{1}{\ell} > u_*$.

THEOREM 5.1. *If the conditions* (5.1)–(5.4) *hold then problem* (P_0) *has at least one solution* (c,u,v) *and* $v'(z) < 0$ *for* $0 < z < \infty$.

The proof is given in the next three sections. In section 6, we solve, for a given $V = V(u)$, (4.23) for (c,Q) under the conditions $Q(u_*) = cu_*$, $Q(1) = 0$. Substituting this solution into (4.24), we solve the resulting equation in section 7 and denote the solution by W; this defines a mapping $T : V \to W$. For technical reasons we shall actually consider the V-equation only in the interval $u_* \le u < 1 - \varepsilon$ ($\varepsilon > 0$), in order to avoid the degeneracy of the leading coefficient at $u = 1$. Thus the mapping T depends on ε (and will be denoted by T_ε), and we shall prove that it has at least one fixed point. In section 8, we let $\varepsilon \to 0$ and, invoking the equivalence established in Lemma 4.2, prove that the solution $(c_\varepsilon, Q_\varepsilon, V_\varepsilon)$ of the fixed point of T_ε converges to a solution of problem (P_1).

6. Solution for (c,Q), given $V(u)$. For any small $\varepsilon > 0$, we introduce the space

$$X_\varepsilon = \{V \in C^0[u_*, 1], \quad 0 \le V \le 1, \quad V \equiv 0 \quad \text{on } [1-\varepsilon, 1]\}.$$

Given $V(u)$ in X_ε, consider the following problem.

Problem (P_2). Find (c,Q) with $c > 0$ such that

$$QQ_u - cQ + d(u, V(u)) f(u, V(u)) = 0 \quad \text{in } \Omega, \tag{6.1}$$

$$Q(u_*) = cu_*, \quad Q(1) = 0, \quad Q > 0 \quad \text{in } \Omega. \tag{6.2}$$

THEOREM 6.1. *Under the assumptions* (5.1), (5.2), *there exists a unique solution to* (6.1), (6.2); *furthermore, the following estimates hold:*

$$\begin{aligned}&c_1 \le c \le c_2, \quad \textit{where}\\ &c_1 = \sqrt{d_0} \max_{\bar\Omega} \left\{ \frac{1}{u} \int_{u_*}^{u} \frac{f(s,0)}{s}\, ds \right\}^{1/2}, \qquad c_2 = 2\sqrt{d_1} \max_{\bar\Omega} \left\{ \frac{1}{u} \int_{u_*}^{u} \frac{f(s,1)}{s}\, ds \right\}^{1/2},\end{aligned} \tag{6.3}$$

$$c_3(1-u) \le Q(u) \le c_2 u \quad \text{in } \Omega, \quad \text{where}$$

$$c_3 = \frac{1}{2}\left\{-c_2 + \left[c_2^2 + 4d_0 \min_{\bar\Omega} \frac{f(s,0)}{1-s}\right]^{1/2}\right\}, \tag{6.4}$$

and

$$Q(u) \le \sqrt{2d_1}\left[\int_u^1 f(s,V(s))ds\right]^{1/2} \le \sqrt{2d_1}||f||_{L^\infty}\sqrt{1-u}. \tag{6.5}$$

Note that by (5.2), $0 < c_1 < c_2$ and $c_3 > 0$.

To prove the theorem, we first consider the following more general problem.

Problem (P_3). Find (c, Q) with $c > 0$ such that

$$QQ_u - cQ + F(u) = 0 \quad \text{in } \Omega = (u_*, 1), \tag{6.6}$$

$$Q(u_*) = cu_*, \quad Q(1) = 0, \quad Q(u) > 0 \quad \text{in } \Omega, \tag{6.7}$$

where

$$F \in C^0[u_*, 1], \quad F(1) = 0, \quad F(u) > 0 \quad \text{in } (u_*, 1). \tag{6.8}$$

LEMMA 6.2. *Problem* (P_3) *admits a unique solution* (c, Q). *The solution has the following properties:*

$$c_F < c \le 2c_F, \quad \text{where } c_F = \max_{\bar\Omega}\left\{\frac{1}{u}\int_{u_*}^u \frac{F(s)}{s}\,ds\right\}^{1/2}, \tag{6.9}$$

$$0 < Q(u) < cu \quad \text{in } \Omega, \tag{6.10}$$

$$\lambda(u)(1-u) \le Q(u) \le \Lambda(u)(1-u) \quad \text{in } \Omega, \tag{6.11}$$

where

$$\lambda(u) = \frac{1}{2}\left\{-c + \left[c^2 + 4\min_{[u,1]} \frac{F(s)}{1-s}\right]^{1/2}\right\} \in (0, \infty),$$

$$\Lambda(u) = \frac{1}{2}\left\{-c + \left[c^2 + 4\max_{[u,1]} \frac{F(s)}{1-s}\right]^{1/2}\right\} \in (0, \infty) \cup \{\infty\},$$

$$Q(u) < \left\{2\int_u^1 F(s)\,ds\right\}^{1/2} \le \left\{2|F|_{L^\infty(\Omega)}\right\}^{1/2}(1-u)^{1/2} \quad \text{in } \Omega \tag{6.12}$$

and

$$c = \int_{u_*}^1 \frac{F(s)}{Q(s)}\,ds > 0. \tag{6.13}$$

Note that if $F_u(1,0) = -\infty$, then $\Lambda(u) = \infty$.

Proof. We first prove uniqueness. Suppose (c_1, Q_1) and (c_2, Q_2) are two solutions. It suffices to show that $c_1 = c_2$. We proceed by contradiction, assuming that $c_1 > c_2$. Then $Q_1(u_*) > Q_2(u_*)$, and we introduce

$$\bar u = \sup\{u \in [u_*, 1], \quad Q_1 > Q_2 \quad \text{in } [u_*, u]\}.$$

If $\bar{u} < 1$, then $Q_1(\bar{u}) = Q_2(\bar{u})$ and $\frac{d}{du}(Q_1 - Q_2) \le 0$ at $\bar{u}$. However, from (6.6), we get

$$\frac{d}{du}(Q_1 - Q_2)|_{u=\bar{u}} = c_1 - c_2 - F(\bar{u})\left[\frac{1}{Q_1(\bar{u})} - \frac{1}{Q_2(\bar{u})}\right] = c_1 - c_2 > 0,$$

a contradiction. We conclude that $\bar{u} = 1$, i.e., $Q_1 > Q_2$ in $[u_*, 1)$. Integrating (6.6) over $[u_*, 1]$ for (c_1, Q_1) and (c_2, Q_2) and taking the difference, we get, after using (6.7),

$$\begin{aligned} 0 > \frac{1}{2}(c_2^2 - c_1^2)u_*^2 &= \left[\frac{1}{2}Q_1^2 - \frac{1}{2}Q_2^2\right]_{u=u_*}^{u=1} = \int_{u_*}^1 \left(Q_1\frac{\partial Q_1}{\partial u} - Q_2\frac{\partial Q_2}{\partial u}\right) du \\ &= \int_{u_*}^1 (c_1 Q_1 - c_2 Q_2)\, du > 0, \end{aligned}$$

a contradiction.

To prove existence we extend $F(u)$ by 0 to $u > 1$. Denote by $Q(c, u)$ $(c > 0)$ the solution to

$$Q_u = c - \frac{F}{Q}\,, \qquad Q(u_*) = cu_* \tag{6.14}$$

and denote by $[u_*, \gamma(c))$ the maximal existence interval where $Q > 0$. The set

$$\mathcal{A} = \{c > 0\,; \quad \gamma(c) > 1\}$$

is clearly an open set. We claim that

$$[0, c_F) \cap \mathcal{A} = \emptyset. \tag{6.15}$$

Indeed, since $F > 0$ in Ω, (6.14) gives $Q_u < c$ so that, by integration, $Q < cu$ and, consequently,

$$Q_u < c - \frac{F}{cu}.$$

Again, by integration,

$$Q < cu - \int_{u_*}^u \frac{F(s)}{cs}\, ds = \frac{u}{c}\left\{c^2 - \frac{1}{u}\int_{u_*}^u \frac{F(s)}{s}\, ds\right\}.$$

It follows that if $c < c_F$, then $\gamma(c) < 1$, i.e., (6.15) holds.

We next prove that

$$(2c_F, \infty) \subset \mathcal{A}. \tag{6.16}$$

Indeed, denote by $\gamma_1(c)$ the smallest value of u in $[u_*, \gamma(c)]$ such that $Q = \frac{1}{2}cu$, i.e.,

$$\gamma_1(c) = \sup\left\{s \le \gamma(c), \quad Q(u) > \frac{1}{2}\, cu \quad \text{in } [u_*, s]\right\}.$$

If $\gamma_1 = \gamma_1(c) \le 1$, then integrating (6.14) over $[u_*, \gamma_1(c)]$, we obtain

$$\begin{aligned} Q(\gamma_1(c)) &= c\gamma_1 - \int_{u_*}^{\gamma_1} \frac{F}{Q} > c\gamma_1 - \int_{u_*}^{\gamma_1} \frac{2F(s)}{cs}\, ds \\ &= \frac{c\gamma_1}{2}\left\{2 - \frac{4}{c^2}\frac{1}{\gamma_1}\int_{u_*}^{\gamma_1} \frac{F(s)}{s}\, ds\right\} > \frac{c\gamma_1}{2} \quad \text{if } c > 2c_F, \end{aligned}$$

which is a contradiction. Hence $\gamma(c) \geq \gamma_1(c) > 1$ if $c \geq 2c_F$ and (6.16) is proved.

Now define $c_0 = \inf\{c;\ c \in \mathcal{A}\}$. By continuity, $\gamma(c_0) \geq 1$ and, from (6.16) and the proof of (6.15), $c_F < c_0 \leq 2c_F$. If $Q(c_0, 1) > 0$, then as $\mathcal{A}$ is open, there are values of c smaller than c_0 (and near c_0) for which $Q(c, u) > 0$ if $u_* \leq u \leq 1$, which contradicts the definition of c_0. We conclude that $Q(c_0, 1) = 0$ and thus $(c_0, Q(c_0, u))$ is a solution to problem (P_3). For simplicity, we shall denote this solution by (c, Q).

Integrating (6.6) over $[u, 1]$, we have

$$\frac{1}{2}[Q^2]_u^1 > -\int_u^1 F(s)\, ds,$$

so that

$$Q^2(u) < 2\int_u^1 F(s)\, ds,$$

and (6.12) follows. Integrating (6.14) over $[u_*, 1]$ and using (6.7), the relation (6.13) also follows.

It remains to prove (6.11). We first derive the upper bound on Q. If we define

$$\delta = \inf\{\tilde{s} > 0; \quad Q(u) < \tilde{s} + \Lambda(u)(1-u) \quad \text{in } [u_*, 1]\},$$

then it suffices to show that $\delta = 0$. Without loss of generality, here we may assume that $\Lambda(1-) < \infty$. Then $\Lambda(u)$ is continuous and, if $\delta > 0$, there exists a point $u_1 \in [u_*, 1)$ such that $Q(u_1) = \delta + \Lambda(u_1)(1 - u_1)$ and

$$\left.\frac{dQ}{du}\right|_{u=u_1} \leq \left.\frac{d}{du}\left[\Lambda(u)(1-u)\right]\right|_{u=u_1} \leq -\Lambda(u_1)$$

since $\Lambda(u)$ is monotone decreasing; the first inequality is actually an equality if $u^* < u_1 < 1$. Using (6.14), we get

$$-\Lambda(u_1) \geq c - \frac{F(u_1)}{\delta + \Lambda(u_1)(1-u_1)} \geq c - \frac{F(u_1)}{\Lambda(u_1)(1-u_1)},$$

that is,

$$-\Lambda^2(u_1) > c\Lambda(u_1) - \frac{F(u_1)}{1-u_1} \geq c\Lambda(u_1) - \max_{[u_1,1]} \frac{F(s)}{1-s}.$$

However, by definition of $\Lambda(u)$,

$$\Lambda^2(u) + c\Lambda(u) - \max_{[u,1]} \frac{F(s)}{1-s} = 0,$$

which is a contradiction.

Similarly, we define

$$\delta = \inf\{s > 0; \quad Q(u) > \lambda(u)(1-u) - s\}$$

and show that $\delta = 0$. □

Proof of Theorem 6.1. Applying Lemma 6.2 with $F(u) = d(u, V(u))f(u, V(u))$ and using (5.1), (5.2), and the fact that $V(u)$ is in X_ε, the assertions of Theorem 6.1 immediately follow. (The assumptions $d_u \leq 0$, $d_v \geq 0$, $ad \leq 1$ are not needed for Theorem 6.1.)

7. A fixed point $(c_\varepsilon, Q_\varepsilon, V_\varepsilon)$. Given $V \in X_\varepsilon$, let (c, Q) denote the solution of (6.1), (6.2) and consider the following problem.

Problem (P_4). Find $W(u)$ such that

$$\frac{Q}{d}\left(\frac{Q}{d}W_u\right)_u - ac\frac{Q}{d}W_u + g(u,W) = 0 \quad \text{in } \Omega_\varepsilon = [u_*, 1-\varepsilon], \tag{7.1}$$

$$-\frac{Q}{d}W_u + \Psi(ac, W) = 0 \quad \text{at } u = u_*, \tag{7.2}$$

$$W(u) = 0 \quad \text{in } [1-\varepsilon, 1], \tag{7.3}$$

where $d = d(u, V(u))$.

LEMMA 7.1. *Under the assumptions* (5.1)–(5.3) *there exists a unique solution* W *to problem* (P_4)*, and the following inequalities hold:*

$$0 \le W \le 1 \quad \textit{in } \Omega_\varepsilon, \tag{7.4}$$

$$-M_\varepsilon \le W_u \le 0 \quad \textit{in } \Omega_\varepsilon, \tag{7.5}$$

$$|W_u(u_*)| \le c_4, \tag{7.6}$$

where M_ε *is a constant which depends on* ε *but not on* $V \in X_\varepsilon$*, and* c_4 *is independent of* ε *and* V*.*

Proof. Extend the definition of $g(u,w)$ to all $w \in \mathbb{R}^1$ such that $g_w < 0$. By Theorem 6.1

$$c_3\varepsilon \le Q < c_2 \quad \text{in } [u_*, 1-\varepsilon] \tag{7.7}$$

so that the differential equation (7.1) is nondegenerate in Ω_ε. The inequalities $g_w < 0$ and $\Psi_w > 0$ will enable us to use the comparison principle for the system (7.1)–(7.3). For any function $\tilde{W}(u)$ satisfying $0 \le \tilde{W}(u) \le 1$, we set $g = g(u, \tilde{W}(u))$, $\Psi = \Psi(ac, \tilde{W}(u))$, and denote the corresponding solution of (7.1)–(7.3) by $W(u)$. Since

$$g(u,0) \ge g(u, \tilde{W}(u)) \ge g(u,1),$$
$$\Psi(ac,0) \le \Psi(ac, \tilde{W}(u)) \le \Psi(ac,1),$$

the comparison principle shows that $0 \le W(u) \le 1$. We can therefore apply the Schauder fixed point argument to the mapping $\tilde{W} \to W$ to deduce the existence of a solution to (7.1)–(7.3). Uniqueness of the solution of (7.1)–(7.3) follows again by a comparison argument.

Next, differentiating (7.1) and setting $Z = \frac{Q}{d}W_u$, we get

$$\left(\frac{Q}{d}Z_u\right)_u - acZ_u + g_v\frac{d}{Q}Z + g_u = 0.$$

Also

$$Z|_{u=u_*} = \Psi(ac, W(u_*)) < 0, \quad Z|_{u=1-\varepsilon} \le 0.$$

Since $g_v < 0$ and $g_u < 0$, the maximum principle yields $Z \le 0$, i.e., $W_u \le 0$. Finally, (7.6) and the first inequality in (7.5) follow from the fact that (7.1) is nondegenerate (i.e., from (7.7)). ∎

We shall now combine Lemma 7.1 with Theorem 6.1. For every $V \in X_\varepsilon$ we define (c, Q) by Theorem 6.1 and W by Lemma 7.1, and introduce the mapping T_ε by

$$T_\varepsilon V = W.$$

Clearly, T_ε maps X_ε into itself, and its image lies in a compact subset of X_ε (since $|W_u| \le M_\varepsilon$). By the uniqueness parts of Theorem 6.1 and Lemma 7.1, it also easily follows that T_ε is continuous. Invoking the Schauder fixed point theorem, we conclude that there exists at least one fixed point for T_ε. We shall denote it by $(c_\varepsilon, Q_\varepsilon, V_\varepsilon)$.

We define a function $u_\varepsilon(z)$ by

$$z = \int_{u_*}^{u_\varepsilon(z)} \frac{d(s, V_\varepsilon(s))}{Q_\varepsilon(s)}\, ds, \quad z \in [0, Z_\varepsilon], \tag{7.8}$$

where

$$Z_\varepsilon = \int_{u_*}^{1-\varepsilon} \frac{d(s, V_\varepsilon(s))}{Q_\varepsilon(s)}\, ds, \tag{7.9}$$

and introduce also the function

$$v_\varepsilon(z) = V_\varepsilon(u_\varepsilon(z)).$$

Then $(c_\varepsilon, u_\varepsilon(z), v_\varepsilon(z))$ form a solution of the system (4.17)–(4.21) in $[0, Z_\varepsilon]$ but without the conditions $u(\infty) = 1$, $v(\infty) = 0$.

By (6.9)–(6.11), the functions $Q_\varepsilon(u)$ are uniformly bounded from above and below by two positive constants, for $u_* \le u \le 1 - \delta$ (for any $\delta > 0$), and the same is true for the constants c_ε. Using also the estimates of Lemma 7.1, we deduce that $u_\varepsilon(0), u'_\varepsilon(0), v_\varepsilon(0), v'_\varepsilon(0)$, are uniformly bounded. Hence we can choose a subsequence $\varepsilon \downarrow 0$ such that

$$(c_\varepsilon, u_\varepsilon, v_\varepsilon) \to (c, u, v), \tag{7.10}$$

where (c, u, v) is a solution of (4.17)–(4.21) for $0 < z < z_0$, where

$$z_0 = \lim_{\varepsilon \downarrow 0} Z_\varepsilon. \tag{7.11}$$

The corresponding limits (Q, V) of $(Q_\varepsilon, V_\varepsilon)$ also exist, and $Q(u) > 0$ if $u_* < u < 1$, $Q(1) = 0$.

LEMMA 7.2. *If $z_0 = \infty$, then (c, u, v) is a solution to problem (P_0).*

Proof. The only assertions that still need to be proved are

$$u(\infty) = 1, \qquad v(\infty) = 0. \tag{7.12}$$

We know that $u(\infty)$, $v(\infty)$ exist. Consequently (by ODE theory) u'' and v'' are bounded and then $u'(\infty)$, $v'(\infty)$ must exist and be equal to zero. The relation (4.30) holds for $0 < z < \infty$, and taking $z \to \infty$, we get $Q(u(\infty)) = 0$ so that $u(\infty) = 1$. Finally, from (4.17), we deduce that

$$d(1, v(\infty))u'' + f(1, v(\infty)) \to 0 \quad \text{if } z \to \infty.$$

From this it follows that $f(1, v(\infty)) = 0$ (otherwise, $u'(z)$ will not converge to zero as $z \to \infty$) so that by (5.2), $v(\infty) = 0$. ∎

8. Proof of Theorem 5.1. In section 7, we established the existence of a fixed point $(c_\varepsilon, u_\varepsilon, v_\varepsilon)$, or $(c_\varepsilon, Q_\varepsilon, V_\varepsilon)$, for the mapping $T_\varepsilon : X_\varepsilon \to X_\varepsilon$ and also proved that $\partial V_\varepsilon/\partial u \le 0$. In this section, we shall replace X_ε by the subset

$$Y_\varepsilon = \Bigg\{ V \in C^0[u_*, 1]; \quad 0 \le V \le 1, \quad V_u \le 0, \\ V \equiv 0 \quad \text{on } [1-\varepsilon, 1], \quad \text{and} \quad V(u) \le \ell(1-u) \quad \text{in } \left[1 - \frac{1}{\ell}, 1\right] \Bigg\}, \tag{8.1}$$

where ℓ is the positive number appearing in condition (5.4).

LEMMA 8.1. *Under the assumptions* (5.1)–(5.4), T_ε *has a fixed point in* Y_ε.

Proof. We need only to show that T_ε maps Y_ε into itself; the rest of the analysis is as before. If we set $T_\varepsilon V = W$, then as before, $W_u \le 0$, and so all we need to show is that

$$W(u) \le \ell(1-u) \quad \text{in } \left[1 - \frac{1}{\ell}, 1 - \varepsilon\right]. \tag{8.2}$$

Consider the function $\bar{W}(u) = \ell(1-u)$ in $[1 - \frac{1}{\ell}, 1 - \varepsilon]$. It satisfies

$$\bar{W}\left(1 - \frac{1}{\ell}\right) = 1 \ge W\left(1 - \frac{1}{\ell}\right), \qquad \bar{W}(1-\varepsilon) = \ell\varepsilon > 0 = W(1-\varepsilon).$$

Also

$$\begin{aligned} \mathcal{L}W &\equiv \frac{Q}{d}\left(\frac{Q}{d}\bar{W}_u\right)_u - \alpha c \frac{Q}{d}\bar{W}_u + g(u, \bar{W}) \\ &= -\ell\left\{\frac{Q}{d}\frac{Q_u}{d} - \frac{Q^2}{d}\frac{d_u + d_v V_u}{d^2}\right\} + \ell a c \frac{Q}{d} + g(u, \bar{W}) \\ &= -\frac{\ell}{d^2}\left\{(c - acd)Q - df\right\} + \frac{Q^2\ell}{d^3}(d_u + d_v V_u) + g(u, \bar{W}) \quad \text{by (6.1)}. \end{aligned}$$

Since $d_u \le 0$, $d_v \ge 0$, $V_u \le 0$, $1 - ad \ge 0$, $Q > 0$, we get

$$\mathcal{L}W \le \ell\left\{\frac{f(u, V(u))}{d(u, V(u))} + \frac{1}{\ell}g(u, \ell(1-u))\right\} \le 0 \quad \text{in } \left[1 - \frac{1}{\ell}, 1 - \varepsilon\right],$$

by (5.4); here we used the inequality $V(u) \le \ell(1-u)$ in $[1 - \frac{1}{\ell}, 1]$. Thus $\bar{W}$ is a supersolution, and (8.2) follows by comparison. ∎

We shall henceforth work with the solution $(c_\varepsilon, u_\varepsilon, v_\varepsilon)$, or $(c_\varepsilon, Q_\varepsilon, V_\varepsilon)$, corresponding to the space Y_ε and with the corresponding limits (c, u, v) (see (7.10)) and (c, Q, V). We shall prove that (c, u, v) is a solution to problem (P_0). In view of Lemma 7.2, all we need to prove is the following.

LEMMA 8.2. *Under the assumptions* (5.1)–(5.4) *the following relations hold:*

$$z_0 = \int_{u_*}^{1} \frac{d(s, V(s))}{Q(s)}\, ds = \infty. \tag{8.3}$$

Proof. Recalling that $f(1, 0) = 0$ and using Lemma 8.1, we have

$$f(u, V_\varepsilon(u)) \le ||f_u||_{L^\infty}(1-u) + ||f_v||_{L^\infty} V_\varepsilon(u) \le M(1-u)$$

for $u \in [1-\frac{1}{\ell}, 1-\varepsilon]$. Using this in (6.5) we get $Q_\varepsilon(u) \le M(1-u)$ with another constant M, independent of ε. It follows that

$$Q(u) \le M(1-u) \quad \text{in } \left[1-\frac{1}{\ell}, 1\right]$$

and, since $d(u, V(u)) \ge d_0 > 0$, the assertion (8.3) follows. □

9. Application to corneal epithelial wound healing. In the model that appears in [2] (see also [1]), after nondimensionalization,

$$\frac{\partial N}{\partial t} = \frac{\partial}{\partial X}\left[(\alpha + 0.1\alpha C)\frac{\partial N}{\partial X}\right] + \frac{\partial}{\partial X}(e(X,t)N) + F(N,C), \tag{9.1}$$

$$\frac{\partial C}{\partial t} = D_C \frac{\partial^2 C}{\partial X^2} + G(N,C), \tag{9.2}$$

where

$$\begin{aligned} F(N,C) &= +(0.9+0.1C)(2N-N^2) - N, \\ G(N,C) &= A + B(N) - \mu N h(C) - \delta C, \\ h(C) &= \frac{C}{\hat{C}+C}, \quad \delta = 242, \\ B(N) &= B(0)\chi_{\{N=0\}}, \quad B(0) > 0, \quad \alpha = 0.012, \quad D_C = 6, \quad \mu = 2\times 10^4. \end{aligned}$$

Here the first term in $F(N,C)$ represents the reduction of the chemical level to equilibrium, and $-N$ is due to the natural death of cells. The function $B(N)$ in $G(N,C)$ is a wound bed source term. The term $h(C)$ in $G(N,C)$ represents chemical reaction between N and C, and $-\delta C$ is a loss due to decay of the chemical stimulus. Finally, $e(X,t)$ accounts for the physiological electric field. This choice is phenomenological, and it is not clear to us what is the physical motivation for a choice of the function $e(X,t)$. On the other hand, if we consider the *effect* of the electric field on the diffusion process, then it is reasonable to assume that it takes the form of a flux $k(N)N_X$, where

$$\begin{aligned} &k(n_*) > 0, \quad k'(N) \le 0 \quad \text{if } N > n_*, \\ &k(N) = 0 \quad \text{if } N > n_* + \varepsilon_0 \end{aligned}$$

for some $\varepsilon_0 > 0$. Indeed, the expression $k(N)N_X$ represents flux, which increases as we get closer to the moving boundary of the healed region, and it becomes negligible at some "distance" away from this boundary (i.e., when $N > n_* + \varepsilon_0$).

In what follows, we replace $e(X,t)N$ by $k(N)N_X$, with $k(N)$ as above, and write

$$k(N) = e_0 E(N), \quad e_0 > 0$$

so that $E'(N) \le 0$; it will be convenient to set $e_0 = \alpha e,\ e > 0$.

Let $X = \sqrt{\alpha}x$, $N = n(x,t)$, $C = 1 + C^* c(x,t)$, where, by [2], $C^* \sim 5$–50. Writing the system (9.1), (9.2) (with $e(X,t)N$ replaced by $\alpha e E(N) N_X$) in the form (1.1), (1.2), we then have

$$d(n,c) = 1.1 + bc + eE(n), \tag{9.3}$$

where $a = \alpha/D_C = 2\times 10^{-3},\ b = 0.1C^*$,

$$f(n,c) = (1+bc)n(1-n) + bnc, \tag{9.4}$$

and

$$g(n,c) = \frac{a}{C^*} G(n, 1 + C^* c).$$

It is easily seen that all the assumptions made in Theorem 1.1 are satisfied.

In order to verify the assumptions of Theorem 5.1, we observe that the numerical results obtained in [1, 2], which exhibit a travelling wave solution, imply that

$$G(1,1) = 0, \qquad G(0, 1 + C^*) = 0$$

so that

$$A = \delta + \mu h(1), \qquad C^* = \frac{1}{\delta}(\mu h(1) + B(0)). \tag{9.5}$$

Setting

$$\theta = \frac{\mu h(1)}{\mu h(1) + B(0)}, \quad 0 < \theta < 1,$$

we get

$$g(n,c) = a\delta \left\{ -c + \theta(1-n) + \theta n \left(1 - \frac{h(1 + C^* c)}{h(1)} \right) + (1-\theta)\chi_{\{n=0\}} \right\}. \tag{9.6}$$

We easily check that the functions d, f, g then satisfy all the assumptions in (5.1)–(5.3). Thus it remains to show that condition (5.4) also holds; this is more dedicate. We note that whereas in [1] (and also in [2]) $\hat{C}$ is taken as 3 so that $h(1) = \frac{1}{4}$, the numerical graph in [1, Figure 2] shows that actually $1 + C^* = 5$ in case $B(0) = 0$ so that, from the second relation in (9.5),

$$h(1) = \frac{\delta C^*}{\mu} \sim \frac{1}{20}.$$

We shall take an intermediate value, $h(1) = \frac{1}{10}$, so that $\hat{C} = 9$. Then

$$g(u,v) = 0.4 \left\{ -v + \theta(1-u) - \theta C^* \frac{uv}{1+bv} \right\}$$

in the healed region. Clearly,

$$\max_{0 \le v \le w} \frac{f(u,v)}{d(u,v)} \le u(1-u) + \frac{buw}{1+bw}$$

and, with $v = \ell(1-u)$,

$$\frac{g(u,v)}{\ell}\Big|_{v=\ell(1-u)} \le -4\theta b \frac{u(1-u)}{1+bw} \quad \text{if } \ell > 1 \ge \theta.$$

Hence the condition (5.4) holds if

$$1 + \frac{\ell b}{1+bw} - \frac{4\theta b}{1+bw} \le 0 \quad \text{for } 0 \le w \le 1. \tag{9.7}$$

This inequality is satisfied for some ℓ, θ near 1 provided $2b > 1$, which is indeed the case since $b = 0.1C^*$ and $C^* > 8$ by (9.5).

We conclude that for the relevant range of parameters that occur in [1, 2] the functions d, f, g satisfy all the assumptions made in Theorem 1.1 and all the assumptions (5.1)–(5.3) made in Theorem 5.1. As for the last assumption, (5.4), made in Theorem 5.1, it does hold for *some* range of relevant parameters.

Acknowledgments. We would like to thank Philip Maini for providing us with the manuscript [2]. We also wish to thank him and E. A. Gaffrey for useful discussions during the preparation of the present paper.

REFERENCES

[1] P. D. Dale, P. K. Maini, and J. A. Sheratt, *Mathematical modeling of corneal epithelial healing*, Math. Biosci., 124 (1994), pp. 127–147.

[2] E. A. Gaffney, P. K. Maini, C. D. McCaig, M. Zhao, and J. V. Forrester, *Modelling corneal epithelial wound closure in the presence of physiological electric fields via a moving boundary formalism*, IMA J. Math. Appl. Med. Biol., 16 (1999), pp. 369–393.

[3] D. A. Ladyzhenskaja, V. A. Solonnikov, and N. N. Ural'ceva, *Linear and Quasilinear Equations of Parabolic Type*, AMS, Providence, RI, 1968.

[4] L. Olsen, P. K. Maini, and J. A. Sheratt, *Spatially varying equilibria of mechanical models: Application to dermal woundn healing*, Math. Biosci., 147 (1998), pp. 113–129.

[5] L. Olsen, P. K. Maini, J. A. Sheratt, and B. Marchant, *Simple modelling of extracellular matrix alignment in dermal wound healing* I: *Cell flux induces alignment*, J. Theor. Med., 1 (1998), pp. 175–192.

[6] L. Olsen, J. A. Sheratt, P. K. Maini, and F. Arnold, *A mathematical model for the capillary endothelial cell-extracellular matrix interactions in wound-healing angiogenesis*, IMA J. Math. Appl. Med. Biol., 14 (1997), pp. 261–281.

[7] L. Olsen, J. A. Sheratt, and P. K. Maini, *A mechanochemical model for adult dermal wound contraction and the permanence for the contracted tissue displacement profile*, J. Theoret. Biol., 177 (1995), pp. 113–128.

SIAM J. MATH. ANAL.
Vol. 32, No. 4, pp. 801–819

INVERSE PROBLEM FOR THE STURM–LIOUVILLE EQUATION ON A SIMPLE GRAPH*

VYACHESLAV PIVOVARCHIK†

Abstract. The spectrum of small vibrations of a graph consisting of three joint smooth strings with the free ends fixed can be reduced to the Sturm–Liouville boundary problem on a graph. This problem occurs also in quantum mechanics. The spectrum of such a problem is investigated in comparison with the union of spectra of Dirichlet problems on the rays of the graph. It is shown that the eigenvalues of the spectra interlace in some sense; thus an analogue of Sturm theorem is established. If the four spectra (the spectrum of the boundary problem on the graph and the three spectra of the mentioned Dirichlet problems) do not intersect, then the inverse problem of recovering the potentials on the rays from the four spectra is uniquely solvable. The procedure of construction of the potentials is presented.

Key words. sinus-type function, function of Hermite–Biehler type, quadratic operator pencil, interpolation series, Dirichlet boundary conditions

AMS subject classifications. 34B24, 34A55, 34B10, 73K03

PII. S0036141000368247

1. Direct problem.

It is well known [5] that if a string is smooth enough (if its density belongs to the Sobolev space W_2^2), then the equation of small transverse vibrations of the string can be reduced by means of Liouville transformation to the Sturm–Liouville equation. We consider the Sturm–Liouville boundary problem on a simple graph that consists of three intervals joined at a common point. The scattering problems on graphs of different forms were considered in many publications in connection with quantum waveguides [1], [2], [4], [7], [8], [14], [16]. The following papers are devoted to inverse problems on graphs: [13], [23], [24]. By inverse problem, we mean recovering the potentials from the known spectra of corresponding boundary problems.

Consider the following boundary problem:

$$y_j'' + \lambda^2 y_j - q_j(x) y_j = 0, \tag{1.1j}$$

$$y_j(\lambda, 0) = 0 \quad (j = 1, 2, 3), \tag{1.2j}$$

$$y_1(\lambda, a) = y_2(\lambda, a) = y_3(\lambda, a), \tag{1.3}$$

$$y_1'(\lambda, a) + y_2'(\lambda, a) + y_3'(\lambda, a) = 0. \tag{1.4}$$

This problem occurs in the following situations: (1) small vibrations of a graph of three inhomogeneous smooth strings each having one end joint and the other one fixed; (2) a quantum particle moving in a quasi-one-dimensional graph domain.

We assume that real-valued functions $q_j(x) \in L_2(0, a)$ $(j = 1, 2, 3)$. For the sake of simplicity let $q_j(x) \geq 0$. Otherwise, we can shift the spectral parameter. We denote by $s_j(\lambda, x)$ $(j = 1, 2, 3)$ the solution of $(1.1j)$ that satisfies the conditions

$$s_j(\lambda, 0) = s_j'(\lambda, 0) - 1 = 0. \tag{1.5j}$$

*Received by the editors February 16, 2000; accepted for publication July 27, 2000; published electronically December 5, 2000.

http://www.siam.org/journals/sima/32-4/36824.html

†Department of Higher Mathematics, Odessa State Academy of Civil Engineering and Architecture, Didrihson Street 4, 65029 Odessa, Ukraine (v.pivorakchik@paco.net).

Then the solutions of equations (1.1j) that satisfy the conditions (1.2j) are

$$y_j(\lambda,x)=C_j s_j(\lambda,x), \tag{1.6j}$$

where C_j are constants. Substituting (1.6j) into (1.3) and (1.4), we obtain the following equation for the eigenvalues of problem (1.1j), (1.2j), (1.3), (1.4):

$$\varphi_1(\lambda)=:\begin{vmatrix} s_1(\lambda,a) & -s_2(\lambda,a) & 0\\ s_1(\lambda,a) & 0 & -s_3(\lambda,a)\\ s_1'(\lambda,a) & s_2'(\lambda,a) & s_3'(\lambda,a)\end{vmatrix}=0$$

or

$$\varphi_1(\lambda)=\psi_1(\lambda)s_3(\lambda,a)+\psi_2(\lambda)s_3'(\lambda,a)=0, \tag{1.7}$$

where

$$\psi_1(\lambda)=:s_1(\lambda,a)s_2'(\lambda,a)+s_2(\lambda,a)s_1'(\lambda,a), \tag{1.8}$$

$$\psi_2(\lambda)=:s_1(\lambda,a)s_2(\lambda,a). \tag{1.9}$$

Let us denote by

$$\varphi_2(\lambda)=:s_1(\lambda,a)s_2(\lambda,a)s_3(\lambda,a), \tag{1.10}$$

by $\{\zeta_k\}_{-\infty,k\ne0}^{\infty}$, the set of zeroes of $\psi_1(\lambda)$, and by $\{\xi_k\}_{-\infty,k\ne0}^{\infty}$ the set of zeroes of $\psi_2(\lambda)$.

LEMMA 1.1. *Under proper enumeration, the following inequalities are valid:*

$$0<\zeta_1<\zeta_2<\cdots<\zeta_k<\cdots,\quad \zeta_{-k}=-\zeta_k.$$

Proof. The set $\{\zeta_k\}_{-\infty,k\ne0}^{\infty}$ coincides with the spectrum of the problem

$$\begin{aligned} y_1''+\lambda^2y_1-q_1(x)y_1&=0,\\ y_2''+\lambda^2y_2-q_2(x)y_2&=0,\\ y_1(\lambda,0)&=y_2(\lambda,0)=0,\\ y_1(\lambda,a)&=y_2(\lambda,a),\\ y_1'(\lambda,a)+y_2'(\lambda,a)&=0,\end{aligned}$$

or, what is the same, with the spectrum of the problem

$$\widetilde{y}''+\lambda^2\widetilde{y}-\widetilde{q}(x)\widetilde{y}=0, \tag{1.11}$$

$$\widetilde{y}(\lambda,0)=\widetilde{y}(\lambda,2a)=0, \tag{1.12}$$

where

$$\widetilde{y}=\begin{cases} y_1(\lambda,x) & \text{if } x\in[0,a],\\ y_2(\lambda,2a-x) & \text{if } x\in(a,2a],\end{cases}$$

$$\widetilde{q}(x)=\begin{cases} q_1(x) & \text{if } x\in[0,a],\\ q_2(2a-x) & \text{if } x\in(a,2a].\end{cases}$$

Thus the set $\{\zeta_k\}_{-\infty,k\ne0}^{\infty}$ coincides with the spectrum of Dirichlet problem (1.11)–(1.12) for the Sturm–Liouville equation with nonnegative potential. The assertion of Lemma 1.1 follows. □

The set $\{\xi_k\}_{-\infty,k\neq 0}^{\infty}$ is the union of the sets of zeroes of the functions $s_1(\lambda,a)$ and $s_2(\lambda,a)$, i.e., the union of the spectra of the following two Dirichlet problems:

$$\begin{aligned} &y_1''+\lambda^2 y_1-q_1(x)y_1=0,\\ &y_1(\lambda,0)=y_1(\lambda,a)=0 \end{aligned} \tag{1.13}$$

and

$$\begin{aligned} &y_2''+\lambda^2 y_2-q_2(x)y_2=0,\\ &y_2(\lambda,0)=y_2(\lambda,a)=0. \end{aligned} \tag{1.14}$$

As usual, we enumerate the zeroes in such a way that $\zeta_{-k}=-\zeta_k$, $\zeta_{k+1}>\zeta_k$, $\xi_{-k}=-\xi_k$, $\xi_{k+1}\geq\xi_k$. Let us introduce the function

$$\Xi(\lambda)=\psi_1(\lambda)+i\alpha\lambda\psi_2(\lambda), \tag{1.15}$$

where an arbitrary constant $\alpha\in(2,\infty)$.

LEMMA 1.2. *The function $\Xi(\lambda)$ can be presented as follows:*

$$\begin{aligned} \Xi(\lambda)&=\frac{\sin 2\lambda a}{\lambda}-\int_0^a (q_1(x)+q_2(x))dx\frac{\cos 2\lambda a}{\lambda^2}+\frac{\upsilon_1(\lambda)}{\lambda^2}\\ &+i\alpha\left(\frac{\sin^2\lambda a}{\lambda}-\int_0^a (q_1(x)+q_2(x))dx\frac{\sin 2\lambda a}{2\lambda^2}+\frac{\sin\lambda a}{\lambda^2}\upsilon_2(\lambda)+\frac{\upsilon_3(\lambda)}{\lambda^3}\right), \end{aligned} \tag{1.16}$$

where $\upsilon_1(\lambda)=-\upsilon_1(-\lambda)\in L_{2a}$, $\upsilon_2(\lambda)=-\upsilon_2(-\lambda)\in L_a$, $\upsilon_3(\lambda)=-\upsilon_3(-\lambda)\in L_{2a}$, and L_a is the class (introduced in [18]) *of entire functions of exponential type $\leq a$ that belong to $L_2(-\infty,\infty)$ for $\lambda\in\mathbb{R}$.*

Proof. We make use of the formulas of [22, section 1.2]:

$$s_j(\lambda,a)=\frac{\sin\lambda a}{\lambda}-\int_0^a q_j(x)dx\frac{\cos\lambda a}{\lambda^2}+\frac{\omega_j(\lambda)}{\lambda^2}\quad(j=1,2,3), \tag{1.17}$$

$$s_j'(\lambda,a)=\cos\lambda a+\int_0^a q_j(x)dx\frac{\sin\lambda a}{\lambda}+\frac{\varrho_j(\lambda)}{\lambda}\quad(j=1,2,3), \tag{1.18}$$

where $\omega_j(\lambda)\in L_a$, $\varrho_j(\lambda)\in L_a$. Substituting (1.17) and (1.18) into (1.8) and (1.9), we obtain

$$\psi_1(\lambda)=\frac{\sin 2\lambda a}{\lambda}-\int_0^a (q_1(x)+q_2(x))dx\frac{\cos 2\lambda a}{\lambda^2}+\frac{\upsilon_1(\lambda)}{\lambda^2}, \tag{1.19}$$

$$\psi_2(\lambda)=\frac{\sin^2\lambda a}{\lambda^2}-\int_0^a (q_1(x)+q_2(x))dx\frac{\sin 2\lambda a}{2\lambda^3}+\frac{\sin\lambda a}{\lambda^3}\upsilon_2(\lambda)+\frac{\upsilon_3(\lambda)}{\lambda^4}, \tag{1.20}$$

where $\upsilon_j(\lambda)$ for $j=1,2,3$ are as described in the statement of Lemma 1.2.

Substituting (1.19) and (1.20) into (1.15), we obtain (1.16). □

Let us denote by $\{\kappa_k\}_{-\infty}^{\infty}$ the set of zeroes of $\Xi(\lambda)$. We enumerate the zeroes in the following way: (1) $\kappa_{-k}=-\overline{\kappa_k}$ for all not pure imaginary κ_k; (2) $\operatorname{Re}\kappa_{k+1}\geq\operatorname{Re}\kappa_k$; (3) the multiplicities are taken into account. We call *proper* this way of enumeration (arbitrary in other respects).

Let us set

$$\Xi_0(\lambda)=:\frac{\sin 2\lambda a}{\lambda}+i\alpha\frac{\sin^2\lambda a}{\lambda}. \tag{1.21}$$

The following enumeration of zeroes $\kappa_k^{(0)}$ of the function $\Xi_0(\lambda)$ is proper:

$$\kappa_{2k-1}^{(0)} = \frac{2\pi k}{a} \quad (k \in N), \tag{1.22}$$

$$\kappa_{2k}^{(0)} = \frac{2\pi k}{a} + \frac{i}{2a}\ln\frac{\alpha+2}{\alpha-2} \quad (k \in N \cup \{0\}), \quad \kappa_k^{(0)} = -\overline{\kappa_k^{(0)}}. \tag{1.23}$$

LEMMA 1.3. *The zeroes of* $\Xi(\lambda)$ *enumerated in the appropriate way behave asymptotically as follows:*

$$\kappa_k = \kappa_k^{(0)} + o(1). \tag{1.24}$$

Proof. Suppose there exists a subsequence $\{\kappa_{k_m}\}_{m=1}^{\infty}$ of the sequence $\{\kappa_k\}_{-\infty}^{\infty}$ such that $\operatorname{Im}\kappa_{k_m} \underset{m\to\infty}{\longrightarrow} \infty$. Then (1.16) implies that

$$\Xi(\kappa_{k_m}) - \frac{e^{-2i\kappa_{k_m}a}}{2i\kappa_{k_m}}\left(1-\frac{\alpha}{2}\right) = o\left(\frac{e^{2|\operatorname{Im}\kappa_{k_m}|a}}{|\kappa_{k_m}|}\right), \tag{1.25}$$

which contradicts the identity $\Xi(\kappa_{k_m}) = 0$. Hence there exists a number $M > 0$ such that $\operatorname{Im}\kappa_k \le M$. In the same way, it can be proved that $\operatorname{Im}\kappa_k$ is bounded below. Hence there exists a constant $M_1 > 0$ such that

$$|\operatorname{Im}\kappa_k| \le M_1.$$

Now it follows from (1.16) and (1.21) that there exists a constant $C > 0$ such that

$$|\Xi(\lambda) - \Xi_0(\lambda)| < \frac{C}{|\lambda|^2}$$

for all $\lambda \in \Pi$, where $\{\lambda : |\operatorname{Im}\lambda| < M + \epsilon\}$. Since the function $\lambda\Xi(\lambda) = \sin 2\lambda a + i\alpha\sin^2\lambda a$ is periodic, for every $r \in (0,\epsilon)$, it is possible to find $d > 0$ such that

$$|\sin 2\lambda a + i\alpha\sin^2\lambda a| > d$$

for all $\lambda \in \Pi\backslash\bigcup_k C_k$, where C_k are circles of radii r with the centers at the points $\kappa_k^{(0)}$. Consequently, for all $\lambda = \{\lambda : \lambda \in \Pi\backslash\bigcup_k C_k, |\lambda| > \frac{C}{d}\}$, the following inequalities are valid:

$$|\Xi_0(\lambda)| > \frac{d}{|\lambda|} > \frac{C}{|\lambda^2|} > |\Xi(\lambda) - \Xi_0(\lambda)|.$$

Since $r > 0$ can be chosen arbitrarily small, we apply the Rouché theorem and obtain the assertion of Lemma 1.3. □

The class of sinus-type functions was introduced in [20].

DEFINITION 1.4. *An entire function* $\omega(\lambda)$ *of exponential type* $\sigma > 0$ *is said to be a function of sinus-type if*

1. *all the zeroes of* $\omega(\lambda)$ *lie in a strip* $|\operatorname{Im}\lambda| < h < \infty$;
2. *for some* h_1 *and all* $\lambda \in \{\lambda : \operatorname{Im}\lambda = h_1\}$, *the following inequalities hold:* $0 < m \le |\omega(\lambda)| \le M < \infty$;
3. *the type of* $\omega(\lambda)$ *in the lower half-plane coincides with that in the upper half-plane.*

LEMMA 1.5.

1. *The function* $\lambda\Xi(\lambda)$ *is of sinus-type.*
2. *The following formula is valid:*

$$\Xi(\lambda) = C\prod_{-\infty}^{\infty}\left(1-\frac{\lambda}{\kappa_k}\right). \tag{1.26}$$

Proof. It follows from Lemma 1.3 that $\lambda\Xi(\lambda)$ satisfies condition (1) of Definition 1.4. This function satisfies condition (2) of Definition 1.4 due to Lemma 1.2. Using (1.16), it is easy to check up that the types of $\lambda\Xi(\lambda)$ in the lower and in the upper half-planes are both equal to $2a$. Assertion (1) of Lemma 1.5 is proved. Now assertion (2) follows [18]. □

The set $\{\kappa_k\}_{-\infty}^{\infty}$ coincides with the spectrum of the problem generated by equations $(1.1j)$ $(j=1,2)$ and boundary conditions $(1.2j)$ $(j=1,2)$ and the following conditions at $x=a$:

$$y_1(\lambda,a) = y_2(\lambda,a), \tag{1.27}$$

$$y_1'(\lambda,a) + y_2'(\lambda,a) + i\alpha\lambda y_2(\lambda,a) = 0. \tag{1.28}$$

This problem admits the following operator interpretation. Denote by A the operator acting in the Hilbert space $H = L_2(0,a)\bigoplus L_2(0,a)\bigoplus\mathbb{C}$ according to the formulas

$$A\begin{pmatrix} y_1(x)\\ y_2(x)\\ y_1(a)\end{pmatrix} = \begin{pmatrix} -y_1''(x)+q_1(x)y_1(x)\\ -y_2''(x)+q_2(x)y_2(x)\\ y_1'(a)+y_2'(a)\end{pmatrix},$$

$$D(A) = \left\{\begin{pmatrix} y_1(x)\\ y_2(x)\\ y_1(a)\end{pmatrix}:\ \begin{matrix} y_1(x)\in W_2^2(0,a), & y_2(x)\in W_2^2(0,a), & y_1(0)=y_2(0)=0,\\ & y_1(a)=y_2(a) & \end{matrix}\right\}.$$

Let us denote by K and P the following operators on the same space:

$$K = \begin{pmatrix} 0&0&0\\ 0&0&0\\ 0&0&\alpha\end{pmatrix}, \qquad P = \begin{pmatrix} I&0&0\\ 0&I&0\\ 0&0&0\end{pmatrix}.$$

We introduce a nonmonic quadratic operator pencil acting in H:

$$L(\lambda) = \lambda^2 P - i\lambda K - A$$

with the domain $D(L(\lambda)) = D(A)$ independent of λ and dense in H.

THEOREM 1.6.

1. *The spectrum* $\{\kappa_k\}_{-\infty}^{\infty}$ *of* $L(\lambda)$ *consists of normal* (*see* [9] *for the definition*) *eigenvalues.*
2. *The geometric multiplicity of each eigenvalue is equal to* 1.
3. *The spectrum is symmetric with respect to the imaginary axis and symmetric eigenvalues possess the same multiplicities.*
4. $\operatorname{Im}\kappa_k \geq 0$ *for all* k.
5. *The point* $\lambda = 0$ *belongs to the resolvent set of* $L(\lambda)$.
6. *All (if any) real eigenvalues are simple.*

Proof. To prove assertion 1, let us notice that the pencil $L(\lambda)$ is a bounded perturbation of the quasi-linear pencil $\lambda^2 I - A$, which has normal spectrum only and it is possible to apply [10, Chapter XI, Theorem 4.2] to the pencil $-A^{-\frac{1}{2}}L(\lambda)A^{-\frac{1}{2}} = I + iA^{-\frac{1}{2}}KA^{-\frac{1}{2}} - \lambda^2 A^{-\frac{1}{2}}PA^{-\frac{1}{2}}$, which has the same spectrum as $L(\lambda)$. Assertion 2 follows from the fact that the problem (1.1j), (1.2j) (j=1,2), (1.27) possesses only one linearly independent solution. Assertion 3 is a consequence of the symmetry of the problem, i.e., of the identity $\Xi(-\bar{\lambda}) = \overline{\Xi(\lambda)}$. We obtain assertion 4 by applying the results of [17] to the pencil $L(-i\lambda)$. Assertion 5 follows from the identity $L(0) = -A$ and from the inequality $A \gg 0$.

Now let us prove assertion 6. Let $\kappa_k \neq 0$ be a real multiple eigenvalue of $L(\lambda)$ and let Y_k and $\dot{Y}_k$ be the corresponding eigen- and associated vectors. Then

$$\kappa_k^2 PY_k - i\kappa_k KY_k - AY_k = 0, \tag{1.29}$$

$$\kappa_k^2 P\dot{Y}_k - i\kappa_k K\dot{Y}_k - A\dot{Y}_k + 2\kappa_k PY_k - iKY_k = 0, \tag{1.30}$$

and consequently (1.29) implies that

$$\kappa_k^2 (PY_k, Y_k) - (AY_k, Y_k) = 0$$

and

$$(KY_k, Y_k) = 0. \tag{1.31}$$

The operator K is nonnegative, and hence

$$KY_k = 0. \tag{1.32}$$

Substituting (1.32) into (1.29), we obtain

$$(\kappa_k^2 P - A)Y_k = 0. \tag{1.33}$$

Equation (1.32) implies that

$$\big(K\dot{Y}_k, Y_k\big) = \big(\dot{Y}_k, KY_k\big) = 0. \tag{1.34}$$

Multiplying (1.30) by Y_k and using (1.31), (1.33), and (1.34), we obtain

$$\begin{aligned} &\kappa_k^2\big(P\dot{Y}_k, Y_k\big) - \big(A\dot{Y}_k, Y_k\big) + 2\kappa_k(PY_k, Y_k) \\ &\quad = \big(\dot{Y}_k, \kappa_k^2 PY_k - AY_k\big) + 2\kappa_k(PY_k, Y_k) = 2\kappa_k(PY_k, Y_k) = 0. \end{aligned} \tag{1.35}$$

The operator P is nonnegative, and therefore (1.35) yields

$$PY_k = 0. \tag{1.36}$$

Equations (1.29), (1.32), and (1.36) imply that

$$AY_k = 0,$$

a contradiction. □

LEMMA 1.7. *A real number κ_k is a zero of $\Xi(\lambda)$ if and only if it is a simple zero of $\psi_1(\lambda)$ and a double zero of $\psi_2(\lambda)$.*

Proof. If κ_k is a real zero of $\Xi(\lambda)$, then definitions (1.8) and (1.9) imply that $\operatorname{Im}\psi_1(\kappa_k) = \operatorname{Im}(\kappa_k\psi_2(\kappa_k)) = 0$. Thus from (1.15), we obtain $\psi_1(\kappa_k) = \kappa_k\psi_2(\kappa_k) = 0$.

Assertion 5 of Theorem 1.6 implies that $\kappa_k \neq 0$. This means that either $s_1(\kappa_k, a) = 0$, or $s_2(\kappa_k, a) = 0$. Let $s_1(\kappa_k, a) = 0$. Then from (1.8), we obtain $s_2(\kappa_k, a)s_1'(\kappa_k, a) = 0$. Since $s_1(\kappa_k, a)$ and $s_1'(\kappa_k, a)$ cannot be equal to zero simultaneous we conclude that $s_2(\kappa_k, a) = 0$. In the same way, $s_2(\kappa_k, a) = 0$ implies that $s_1(\kappa_k, a) = 0$. Also, it is clear from (1.15) that if $\psi_1(\kappa_k) = \psi_2(\kappa_k) = 0$, then $\Xi(\kappa_k) = 0$. □

Remark 1.1. The direct and inverse problems for the operator pencil $L(\lambda)$ are considered in details in [26].

DEFINITION 1.8. *An entire function $\omega(\lambda)$ is said to be of Hermite–Biehler (HB) class if it has no zeroes in the closed lower half-plane* $\operatorname{Im}\lambda \leq 0$ *and*

$$\left|\frac{\omega(\lambda)}{\bar{\omega}(\lambda)}\right| < 1 \quad \textit{for } \operatorname{Im}\lambda > 0.$$

Here by $\bar{\omega}(\lambda)$, we mean the entire function obtained from $\omega(\lambda)$ by replacing the coefficients in its Taylor series by their complex-conjugates.

DEFINITION 1.9 (see [18, p. 313]). *An entire function $\omega(\lambda)$ that has no zeroes in the open lower half-plane and satisfies the condition*

$$\left|\frac{\omega(\lambda)}{\bar{\omega}(\lambda)}\right| \leq 1 \quad \textit{for } \operatorname{Im}\lambda > 0 \tag{1.37}$$

is said to be a function of generalized Hermite–Biehler ($\overline{HB}$) class.

Let us arrange the sequence $\{\kappa_k\}_{-\infty}^{\infty}$ into two subsequences $\{\kappa_{k_p}\}_{p=-\infty}^{\infty}$ and $\{\kappa_{k_j}\}_{j=-n,j\neq 0}^{n}$ ($k_{-j} = -k_j$ and $k_{-p} = -k_p$ and, consequently, $\kappa_{k_{-j}} = -\bar{\kappa}_{k_j}$ and $\kappa_{k_{-p}} = -\bar{\kappa}_{k_p}$) such that $\{\kappa_{k_p}\}_{p=-\infty}^{\infty} \bigcup \{\kappa_{k_j}\}_{j=-n,j\neq 0}^{n} = \{\kappa_k\}_{-\infty}^{\infty}$, $n \leq \infty$, $\operatorname{Im}\kappa_{k_j} = 0$ for all j and $\operatorname{Im}\kappa_{k_p} > 0$ for all p. Now we can rewrite (1.26) as follows:

$$\Xi(\lambda) = C \prod_{j=-n,j\neq 0}^{n} \left(1 - \frac{\lambda}{\kappa_{k_j}}\right) \tilde{\Xi}(\lambda), \tag{1.38}$$

where

$$\tilde{\Xi}(\lambda) =: \prod_{p=-\infty}^{\infty} \left(1 - \frac{\lambda}{\kappa_{k_p}}\right). \tag{1.39}$$

LEMMA 1.10.

$$\tilde{\Xi}(\lambda) \in HB. \tag{1.40}$$

Proof. The following inequality is a consequence of (1.22)–(1.24):

$$\sum_{p=-\infty}^{\infty} \left|\operatorname{Im}\frac{1}{\kappa_{k_p}}\right| < \infty.$$

Now the assertion of the lemma follows from M.G. Krein's theorem [18, section VII.3, Theorem 6]. □

COROLLARY 1.11.

$$\Xi(\lambda) \in \overline{HB}. \tag{1.41}$$

Proof. Evidently,

$$\left|1-\frac{\lambda}{\kappa_{k_p}}\right| \le \left|1-\frac{\lambda}{\bar\kappa_{k_p}}\right|$$

for all $p=0,\pm1,\pm2,\dots$ and all λ from the open upper half-plane. This implies that

$$\left|\frac{\Xi(\lambda)}{\bar\Xi(\lambda)}\right| = \prod_{-\infty}^{\infty}\left|1-\frac{\lambda}{\kappa_{k_p}}\right|\left|1-\frac{\lambda}{\bar\kappa_{k_p}}\right|^{-1} \le 1,$$

and (1.41) follows. □

COROLLARY 1.12. *The sequences* $\{\zeta_k\}_{-\infty,k\ne0}^{\infty}$ *and* $\{\xi_k\}_{-\infty,k\ne0}^{\infty}\cup\{0\}$ *interlace in the usual sense:*

$$0\le\zeta_1\le\xi_1\le\cdots\le\zeta_k\le\xi_k\le\cdots\quad(\zeta_{-k}=-\zeta_k)\quad(\xi_{-k}=-\xi_k).$$

Proof. This corollary follows from [18, Chapter 7.2, Theorem 3′]. □

THEOREM 1.13. *The sequences* $\{\zeta_k\}_{-\infty,k\ne0}^{\infty}$ *and* $\{\xi_k\}_{-\infty,k\ne0}^{\infty}\cup\{0\}$ *interlace in the following sense:*

1. $0<\zeta_1<\xi_1$;
2. *for every* $n>1$ *the following alternative is valid: either the interval* (ζ_1,ζ_n) *contains exactly* $n-1$ *(counting multiplicities) elements of the set* $\{\xi_k\}_1^{\infty}$*, and then* $\zeta_n\notin\{\xi_k\}_1^{\infty}$*, or the interval* (ζ_1,ζ_n) *contains exactly* $n-2$ *(counting multiplicities) elements of the set* $\{\xi_k\}_1^{\infty}$*, and then* $\zeta_n\in\{\xi_k\}_1^{\infty}$*. Here* $\zeta_{-k}=-\zeta_k$ *and* $\xi_{-k}=-\xi_k$.

Proof. Since the function $\tilde\Xi(\lambda)\in HB$, the zeroes $\{\zeta_{k_p}\}_{-\infty,p\ne0}^{\infty}$ of the function $\frac{\tilde\Xi(\lambda)+\tilde\Xi(-\lambda)}{2}$ and the zeroes $\{\xi_{k_p}\}_{-\infty,p\ne0}^{\infty}$ of the function $\frac{\tilde\Xi(\lambda)-\tilde\Xi(-\lambda)}{2i\lambda}$ interlace in the strict sense due to N. Meiman's theorem (see [18, Chapter 7.2, Theorem 3]):

$$0<\zeta_{k_1}<\xi_{k_1}<\cdots<\zeta_{k_j}<\xi_{k_j}<\cdots\quad(\zeta_{k_{-1}}=-\zeta_{k_1})\quad(\xi_{k_{-1}}=-\xi_{k_1}).\tag{1.42}$$

Combining (1.42) with Lemma 1.7 and taking into account Corollary 1.12, we finish the proof. □

Let us denote by $\{\tau_k\}$ the set of zeroes of the function $\varphi_1(\lambda)$ (see (1.7) for the definition) and by $\{\theta_k\}$ the set of zeroes of the function $\varphi_2(\lambda)$ (see (1.10) for the definition).

LEMMA 1.14. *All* τ_k *and* θ_k *are real and nonzero.*

Proof. Consider the operator $\tilde A$ acting in $L_2(0,a)\bigoplus L_2(0,a)\bigoplus L_2(0,a)\bigoplus\mathbb{C}$:

$$\tilde A\begin{pmatrix}y_1(x)\\y_2(x)\\y_3(x)\\y_1(a)\end{pmatrix}=\begin{pmatrix}-y_1''(x)+q_1(x)y_1(x)\\-y_2''(x)+q_2(x)y_2(x)\\-y_3''(x)+q_3(x)y_3(x)\\y_1'(a)+y_2'(a)+y_3'(a)\end{pmatrix},$$

$$D(\tilde A)=\left\{\begin{pmatrix}y_1(x)\\y_2(x)\\y_3(x)\\y_1(a)\end{pmatrix}:\ \begin{array}{c}y_j(x)\in W_2^2(0,a),\quad y_j(0)=0\quad\text{for }(j=1,2,3),\\ y_1(a)=y_2(a)=y_3(a)\end{array}\right\}.$$

It is easy to check up that this operator is self-adjoint, and due to $q_j(x)\ge0$ for $j=1,2,3$, we obtain $\tilde A\gg0$, i.e., $\tilde A\ge\epsilon I$, where $\epsilon>0$. The set $\{\tau_k^2\}_1^{\infty}$ coincides with

the spectrum of $\tilde{A}$, and consequently $\tau_k^2 > 0$ for all k. Here we enumerate τ_k and θ_k in such a way that $\tau_{-k} = -\tau_k$ and $\theta_{-k} = -\theta_k$. The set $\{\theta_k\}_{-\infty,k\neq 0}^{\infty}$ coincides with the union of the spectra of the following three Dirichlet problems:

$$y_j'' + \lambda^2 y_j - q_j(x) y_j = 0, \tag{1.1j}$$

$$y_j(\lambda, 0) = y_j(\lambda, a) = 0 \quad (j = 1, 2, 3). \tag{1.43j}$$

The assertions of the lemma follow. □

Now it is possible to enumerate the sets $\{\tau_k\}_{-\infty,k\neq 0}^{\infty}$ and $\{\theta_k\}_{-\infty,k\neq 0}^{\infty}$ in the usual way: $\tau_{-k} = -\tau_k$, $\tau_k \le \tau_{k+1}$, and $\theta_{-k} = -\theta_k$, $\theta_k \le \theta_{k+1}$.

LEMMA 1.15.

1. *If $\tau_k = \theta_n$ for some k and n, then $\frac{d\varphi_2(\lambda)}{d\lambda}|_{\lambda=\theta_n} = 0$, i.e., at least two of the three functions $s_1(\lambda, a)$, $s_2(\lambda, a)$, $s_3(\lambda, a)$ have (simple) zeroes at $\lambda = \theta_n = \tau_k$.*
2. *If $\tau_k = \theta_n$ and $\frac{d\varphi_1(\lambda)}{d\lambda}|_{\lambda=\theta_n} = 0$, then $s_1(\theta_n, a) = s_2(\theta_n, a) = s_3(\theta_n, a) = 0$ and $\frac{d^2\varphi_1(\lambda)}{d\lambda^2}|_{\lambda=\theta_n} \neq 0$.*

Proof. By definition of θ_n we have $\varphi_2(\theta_n) = 0$ and hence definition (1.10) implies that $s_j(\theta_n, a) = 0$ for at least one j $(j = 1, 2, 3)$, say, $s_3(\theta_n, a) = 0$. Then equations $\varphi_1(\theta_n) = \varphi_1(\tau_k) = 0$ and (1.7) imply that $\psi_2(\theta_n) = 0$ and assertion 1 of Lemma 1.15 follows. If we assume not $s_3(\theta_n, a) = 0$ but $s_2(\theta_n, a) = 0$ or $s_1(\theta_n, a) = 0$, then the proof is analogous.

Now let $\frac{d\varphi_1(\lambda)}{d\lambda}|_{\lambda=\theta_n} = 0$; then

$$\begin{aligned} \left.\frac{d\varphi_1(\lambda)}{d\lambda}\right|_{\lambda=\theta_n} &= \left.\frac{d\psi_1(\lambda)}{d\lambda}\right|_{\lambda=\theta_n} s_3(\theta_n, a) + \psi_1(\theta_n) \left.\frac{\partial s_3(\lambda, a)}{\partial\lambda}\right|_{\lambda=\theta_n} \\ &\quad + \left.\frac{d\psi_2(\lambda)}{d\lambda}\right|_{\lambda=\theta_n} s_3'(\theta_n, a) + \psi_2(\theta_n) \left.\frac{\partial s_3'(\lambda, a)}{\partial\lambda}\right|_{\lambda=\theta_n} = 0. \end{aligned} \tag{1.44}$$

According to assertion 1, at least two of the functions $s_j(\lambda, a)$ have simple zeros at $\lambda = \theta_n$. Let $s_1(\theta_n, a) = s_2(\theta_n, a) = 0$ and, consequently, $\frac{\partial s_1(\lambda,a)}{\partial\lambda}|_{\lambda=\theta_n} \neq 0$ and $\frac{\partial s_2(\lambda,a)}{\partial\lambda}|_{\lambda=\theta_n} \neq 0$. Then from (1.44), we obtain

$$\begin{aligned} &s_3(\theta_n, a) \left(s_2'(\theta_n, a) \left.\frac{ds_1(\lambda, a)}{d\lambda}\right|_{\lambda=\theta_n} + s_1'(\theta_n, a) \left.\frac{ds_2(\lambda, a)}{d\lambda}\right|_{\lambda=\theta_n} \right) \\ &\quad = s_3(\theta_n, a) \left.\frac{d}{d\lambda}(s_2(\lambda, a) s_1'(\lambda, a) + s_1(\lambda, a) s_2'(\lambda, a))\right|_{\lambda=\theta_n} \\ &\quad = s_3(\theta, a) \left.\frac{d\psi_1(\lambda)}{d\lambda}\right|_{\lambda=\theta_n} = 0. \end{aligned} \tag{1.45}$$

Since all zeroes of the function $\psi_1(\lambda)$ are simple (see Lemma 1.1), (1.45) implies that $s_3(\theta_n) = 0$. It is clear that the geometric multiplicity of each eigenvalue of problem (1.1j), (1.2j), (1.3), (1.4) is not more then 2, which means that $\frac{d^2\varphi_1(\lambda)}{d\lambda^2}|_{\lambda=\theta_n} \neq 0$. If we assume that $s_1(\theta_n, a) = s_2(\theta_n, a) = 0$, then the proof is analogous. □

LEMMA 1.16. *The function $\varphi_1(\lambda)$ can be presented as follows:*

$$\begin{aligned} \varphi_1(\lambda) &= 3\frac{\sin^2\lambda a}{\lambda^2}\cos\lambda a - 2\frac{\sin\lambda a\cos^2\lambda a}{\lambda^3}(B_1 + B_2 + B_3) \\ &\quad + \frac{\sin^3\lambda a}{\lambda^3}(B_1 + B_2 + B_3) + (B_1B_2 + B_2B_3 + B_1B_3)\frac{\cos^3\lambda a}{\lambda^4} \\ &\quad + \omega_1(\lambda)\frac{\sin\lambda a}{\lambda^3} + \frac{\omega_2(\lambda)}{\lambda^4}, \end{aligned} \tag{1.46}$$

where $\omega_1(\lambda) \in L_{2a}$, $\omega_2(\lambda) \in L_{3a}$ *and* $B_j = \int_0^a q_j(x)dx$.

Proof. To prove this lemma it is sufficient to substitute (1.17) and (1.18) into (1.7) and to make use of (1.8) and (1.9). □

LEMMA 1.17. *The set* $\{\tau_k\}_{-\infty,k\neq0}^{\infty}$ *of zeroes of* $\varphi_1(\lambda)$ *can be presented as the union of three subsequences* $\{\rho_k^{(1)}\}_{-\infty,k\neq0}^{\infty} \cup \{\rho_k^{(2)}\}_{-\infty,k\neq0}^{\infty} \cup \{\rho_k^{(3)}\}_{-\infty,k\neq0}^{\infty}$, *which, being enumerated in the usual way* ($\rho_k^{(j)} = -\rho_{-k}^{(j)}$, $\rho_k \le \rho_{k+1}$), *behave asymptotically as follows:*

$$\rho_k^{(1)} \underset{k\to\infty}{=} \frac{\pi k}{a} + \frac{M_1}{k} + \frac{\beta_k^{(1)}}{k}, \tag{1.47}$$

$$\rho_k^{(2)} \underset{k\to\infty}{=} \frac{\pi k}{a} + \frac{M_2}{k} + \frac{\beta_k^{(2)}}{k}, \tag{1.48}$$

$$\rho_k^{(3)} \underset{k\to\infty}{=} \frac{\pi(k-\frac{1}{2})}{a} + \frac{1}{3k}(B_1+B_2+B_3) + \frac{\beta_k^{(3)}}{k}, \tag{1.49}$$

where $\{\beta_k^{(j)}\}_{-\infty,k\neq0}^{\infty} \in l_2$ *for* $j = 1,2,3$ *and* M_1 *and* M_2 *are the solutions (both real and may be equal) of the equation*

$$3M^2 - 2\pi^{-1}(B_1+B_2+B_3)M + \pi^{-2}(B_1B_2+B_1B_3+B_2B_3) = 0. \tag{1.50}$$

Proof. It can be proved in the same way as Lemma 1.3 that the set of zeroes $\{\tau_k\}_{-\infty,k\neq0}^{\infty}$ can be arranged in the following way: $\{\tau_k\}_{-\infty,k\neq0}^{\infty} = \{\rho_k^{(1)}\}_{-\infty,k\neq0}^{\infty} \cup \{\rho_k^{(2)}\}_{-\infty,k\neq0}^{\infty} \cup \{\rho_k^{(3)}\}_{-\infty,k\neq0}^{\infty}$, where

$$\rho_k^{(1)} \underset{k\to\infty}{=} \frac{\pi k}{a} + o(1), \tag{1.51}$$

$$\rho_k^{(2)} \underset{k\to\infty}{=} \frac{\pi k}{a} + o(1), \tag{1.52}$$

$$\rho_k^{(3)} \underset{k\to\infty}{=} \frac{\pi(k-\frac{1}{2})}{a} + o(1). \tag{1.53}$$

Substituting (1.51)–(1.53) into the equation

$$\varphi_1(\rho_k^{(j)}) = 0,$$

where $\varphi_1(\lambda)$ is given by (1.46) we expand the right-hand side of the resulting equation in power series. Then using the formulas $\{\omega_j(\rho_k^{(i)})\}_{-\infty,k\neq0}^{\infty} \in l_2$ for $j = 1,2; i = 1,2,3$ (see [22, Lemma 1.4.3]), we obtain (1.47)–(1.49). □

Let us introduce the function

$$\Upsilon(\lambda) = \varphi_1(\lambda) + i\beta\lambda\varphi_2(\lambda), \tag{1.54}$$

where $\beta \in (3,\infty)$ is an arbitrary constant.

LEMMA 1.18. *The function* $\Upsilon(\lambda)$ *can be presented as follows:*

$$\Upsilon(\lambda) = 3\frac{\sin^2\lambda a}{\lambda^2}\cos\lambda a + i\beta\frac{\sin^3\lambda a}{\lambda^2} - \frac{\sin\lambda a\cos 2\lambda a}{\lambda^3}\int_0^a (q_1(x)+q_2(x)+q_3(x))\,dx$$
$$+ -i\beta\frac{\sin^2\lambda a\cos\lambda a}{\lambda^3}\int_0^a (q_1(x)+q_2(x)+q_3(x))dx + \frac{\upsilon_3(\lambda)}{\lambda^3}, \tag{1.55}$$

where $v_3(\lambda) \in L_{3a}$.

Proof. To prove this statement it is enough to substitute (1.7)–(1.10) into (1.54) and to make use of (1.17)–(1.20). □

Let us introduce the following function:

$$\Upsilon^{(0)}(\lambda) = 3\frac{\sin^2 \lambda a}{\lambda^2}\cos \lambda a + i\beta\frac{\sin^3 \lambda a}{\lambda^2}. \tag{1.56}$$

Let us denote by $\{\lambda_k\}_{-\infty}^{\infty}$ the set of zeroes of $\Upsilon(\lambda)$ and by $\{\lambda_k^{(0)}\}_{-\infty}^{\infty}$ that of $\Upsilon_0(\lambda)$. Under proper enumeration the zeroes $\{\lambda_k^{(0)}\}_{-\infty}^{\infty}$ can be arranged into three subsequences:

$$\lambda_{3k}^{(0)} = \frac{\pi k}{a} + \frac{i}{a}\log\left(\frac{\beta+3}{\beta-3}\right), \quad k \in \mathbb{N} \cup \{0\}, \tag{1.57}$$

$$\lambda_{3k-1}^{(0)} = \lambda_{3k-2}^{(0)} = \frac{\pi k}{a}, \quad k \in \mathbb{N}. \tag{1.58}$$

Here $\lambda_{-k}^{(0)} = -\overline{\lambda_k^{(0)}}$ for all $k \neq 0$.

LEMMA 1.19. *The zeroes* $\{\lambda_k\}_{-\infty}^{\infty}$ *behave asymptotically as follows:*

$$\lambda_k \underset{k\to\infty}{=} \lambda_k^{(0)} + o(1). \tag{1.59}$$

Proof. The proof of this lemma is analogous to that of Lemma 1.3. □

LEMMA 1.20.

1. *The function* $\lambda^2\Upsilon(\lambda)$ *is of sinus-type.*
2. *The following formula is valid:*

$$\Upsilon(\lambda) = C\prod_{-\infty}^{\infty}\left(1 - \frac{\lambda}{\lambda_k}\right), \tag{1.60}$$

where C *is a constant.*

Proof. It follows from Lemma 1.19 and from formulas (1.57), (1.58) that $\lambda^2\Upsilon(\lambda)$ satisfies condition (1) of Definition 1.4. This function satisfies also condition (2) of Definition 1.4 due to Lemma 1.18. Using (1.55) it is easy to check up that the types of $\lambda^2\Upsilon(\lambda)$ in the lower and in the upper half-planes are equal to $3a$ both. Assertion (1) of Lemma 1.20 is proved. Now the assertion (2) follows [20]. □

Let us introduce the operators acting in $L_2(0,a) \bigoplus L_2(0,a) \bigoplus L_2(0,a) \bigoplus \mathbb{C}$:

$$\tilde{K} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \beta \end{pmatrix}, \qquad \tilde{P} = \begin{pmatrix} I & 0 & 0 & 0 \\ 0 & I & 0 & 0 \\ 0 & 0 & I & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

It is clear that $\tilde{A} = \tilde{A}^* \gg 0$, $\tilde{P} \geq 0$, $\tilde{K} \geq 0$. We consider a nonmonic quadratic operator pencil of the form

$$\tilde{L}(\lambda) = \lambda^2\tilde{P} - i\lambda\tilde{K} - \tilde{A}$$

with the domain $D(\tilde{L}(\lambda)) = D(\tilde{A})$. We identify the spectrum $\{\lambda_k\}_{-\infty}^{\infty}$ of $\tilde{L}(\lambda)$ with the spectrum of the problem generated by (1.1j), (1.2j) (j=1,2,3), (1.3) and by the following boundary condition:

$$y_1'(\lambda,a) + y_2'(\lambda,a) + y_3'(\lambda,a) + i\beta\lambda y_1(\lambda,a) = 0. \tag{1.61}$$

This problem has the following physical sense. It describes small vibration of the mentioned graph of three strings damped at the point of connection.

THEOREM 1.21.

1. *The spectrum* $\{\lambda_k\}_{-\infty}^{\infty}$ *consists of normal eigenvalues.*
2. *The geometric multiplicity of each eigenvalue is* ≤ 2.
3. *The spectrum is symmetric with respect to the imaginary axis and symmetric points have the same multiplicities.*
4. $\operatorname{Im}\lambda_k \ge 0$.
5. *The point* $\lambda = 0$ *belongs to the resolvent set of* $\tilde{L}(\lambda)$.
6. *All (if any) real eigenvalues do not possess associated eigenvectors.*

Proof. Assertion 1 is true because the spectrum of $\tilde{L}(\lambda)$ coincides with the set of zeroes of $\Upsilon(\lambda)$, which is an entire function. Assertion 2 follows from the fact that problem (1.1j), (1.2j), ($j = 1, 2, 3$), (1.3) possesses exactly two linearly independent solutions. Assertion 3 is a consequence of the symmetry of the problem. We obtain assertion 4 if we apply the results of [17] to the pencil $\tilde{A}^{-\frac{1}{2}}\tilde{L}(-i\lambda)\tilde{A}^{-\frac{1}{2}}$. Assertion 5 follows from the inequality $\tilde{A} \gg 0$. The proof of assertion 6 is quite the same as that of assertion 6 of Theorem 1.6. □

Let us rearrange the sequence $\{\lambda_k\}_{-\infty}^{\infty}$ into two subsequences $\{\lambda_{k_p}\}_{p=-\infty}^{\infty}$ and $\{\lambda_{k_j}\}_{j=-n,j\ne 0}^{n}$, ($k_{-j} = -k_j$ and $k_{-p} = -k_p$ and, consequently, $\lambda_{k_{-j}} = -\bar{\lambda}_{k_j}$ and $\lambda_{k_{-p}} = -\bar{\lambda}_{k_p}$) such that $\{\lambda_{k_p}\}_{p=-\infty}^{\infty} \bigcup \{\lambda_{k_j}\}_{j=-n,j\ne 0}^{n} = \{\lambda_k\}_{-\infty}^{\infty}$, $n \le \infty$, $\operatorname{Im}\lambda_{k_j} = 0$ for all j and $\operatorname{Im}\lambda_{k_p} > 0$ for all p. Now we can rewrite (1.60) as follows:

$$\Upsilon(\lambda) = C\prod_{j=-n,j\ne 0}^{n}\left(1-\frac{\lambda}{\lambda_{k_j}}\right)\tilde{\Upsilon}(\lambda), \tag{1.62}$$

where

$$\tilde{\Upsilon}(\lambda) =: \prod_{p=-\infty}^{\infty}\left(1-\frac{\lambda}{\lambda_{k_p}}\right). \tag{1.63}$$

LEMMA 1.22.

$$\tilde{\Upsilon}(\lambda) \in HB. \tag{1.64}$$

The proof of this lemma is quite the same as that of Lemma 1.10.

COROLLARY 1.23.

$$\Upsilon(\lambda) \in \overline{HB}. \tag{1.65}$$

The proof of this corollary is quite the same as that of Corollary 1.11.

COROLLARY 1.24. *The sequences* $\{\tau_k\}_{-\infty,k\ne 0}^{\infty}$ *and* $\{\theta_k\}_{-\infty,k\ne 0}^{\infty} \cup \{0\}$ *interlace in the usual sense:*

$$0 \le \tau_1 \le \theta_1 \le \cdots \le \tau_k \le \theta_k \le \cdots \quad (\tau_{-k} = -\tau_k) \quad (\theta_{-k} = -\theta_k).$$

Proof. This corollary follows from [18, Chapter 7.2, Theorem 3′]. □

THEOREM 1.25. *The sets* $\{\tau_k\}_{-\infty,k\ne 0}^{\infty}$ *and* $\{\theta_k\}_{-\infty}^{\infty}$ *(we set* $\theta_0 = 0$*) interlace in the following sense:*

1. $\theta_0 < \tau_1 < \theta_1$.

2. *For each simple* τ_n $(n > 1)$, *either*

$$\theta_{n-1} < \tau_n < \theta_n$$

or

$$\tau_{n-1} < \theta_{n-1} = \tau_n = \theta_n < \tau_{n+1}.$$

3. *For each double* $\tau_n = \tau_{n+1}$ $(n > 1)$,

$$\tau_{n-1} < \theta_{n-1} = \tau_n = \theta_n = \tau_{n+1} = \theta_{n+1} < \tau_{n+2}.$$

4. *The multiplicity of each* τ_n *is* ≤ 2.

Proof. The function $\tilde{\Upsilon}(\lambda) \in HB$, and consequently the zeroes $\{\tau_{k_p}\}_{-\infty, p\neq 0}^{\infty}$ of the function $\frac{\tilde{\Upsilon}(\lambda)+\tilde{\Upsilon}(-\lambda)}{2}$ and the zeroes $\{\theta_{k_p}\}_{-\infty, p\neq 0}^{\infty}$ of the function $\frac{\tilde{\Xi}(\lambda)-\tilde{\Xi}(-\lambda)}{2i\lambda}$ interlace in the usual sense due to N. Meiman's theorem (see [18, Chapter 7.2, Theorem 3]:

$$0 < \tau_{k_1} < \theta_{k_1} < \cdots < \tau_{k_j} < \theta_{k_j} < \cdots \quad (\tau_{k_{-1}} = -\tau_{k_1}) \quad (\theta_{k_{-1}} = -\theta_{k_1}). \tag{1.66}$$

Combining (1.66) with Lemma 1.15 and taking into account Corollary 1.24, we finish the proof. □

2. Inverse problem. Here we deal with the problem of recovering of the potentials $\{q_1(x), q_2(x), q_3(x)\}$ from the spectral data. Denote by Q the set of triplets $\{q_1(x), q_2(x), q_3(x)\}$, which satisfy the following conditions: the real-valued functions $q_j(x) \in L_2(0, a)$ $(j = 1, 2, 3)$.

THEOREM 2.1. *Let the following conditions be valid:*

1. *Three sequences* $\{\nu_k^{(j)}\}_{-\infty, k\neq 0}^{\infty}$ $(j = 1, 2, 3)$ *of real numbers are such that*
   (i) $\nu_{-k}^{(j)} = -\nu_k^{(j)}$, $\nu_k^{(j)} \neq 0$ *for all* k *and* j;
   (ii) $\{\nu_k^{(1)}\}_{-\infty, k\neq 0}^{\infty} \bigcap \{\nu_k^{(2)}\}_{-\infty, k\neq 0}^{\infty} = \emptyset$; $\{\nu_k^{(2)}\}_{-\infty, k\neq 0}^{\infty} \bigcap \{\nu_k^{(3)}\}_{-\infty, k\neq 0}^{\infty} = \emptyset$; $\{\nu_k^{(1)}\}_{-\infty, k\neq 0}^{\infty} \bigcap \{\nu_k^{(3)}\}_{-\infty, k\neq 0}^{\infty} = \emptyset$;
   (iii)

$$\nu_k^{(j)} \underset{k\to\infty}{=} \frac{\pi k}{a} + \frac{B_j}{k} + \frac{\delta_k^{(j)}}{k^2}, \tag{2.1}$$

   where B_j *are real constants,* $B_1 \neq B_2$, $B_2 \neq B_3$, $B_1 \neq B_3$, $\{\delta_k^{(j)}\}_{-\infty, k\neq 0} \in l_2$ *for* $(j = 1, 2, 3)$.
2. *A sequence* $\{\tau_k\}_{-\infty, k\neq 0}^{\infty}$ *of real numbers* $(\tau_{-k} = -\tau_k,\ \tau_k \le \tau_{k+1})$ *can be presented as a union of three subsequences* $\{\tau_k\}_{-\infty, k\neq 0}^{\infty} = \{\rho_k^{(1)}\}_{-\infty, k\neq 0}^{\infty} \cup \{\rho_k^{(2)}\}_{-\infty, k\neq 0}^{\infty} \cup \{\rho_k^{(3)}\}_{-\infty, k\neq 0}^{\infty}$ $(\rho_{-k}^{(j)} = -\rho_k^{(j)}$ *and* $\rho_k^{(j)} < \rho_{k+1}^{(j)})$ *that behave asymptotically as follows:*

$$\rho_k^{(j)} \underset{k\to\infty}{=} \frac{\pi k}{a} - \frac{M_j}{k} + \frac{\beta_k^{(j)}}{k^2}, \quad k \in \mathbb{N}, \quad j = 1, 2; \tag{2.2}$$

$$\rho_k^{(3)} \underset{k\to\infty}{=} \frac{\pi(k - \frac{1}{2})}{a} + \frac{B_0}{3k} + \frac{\beta_k^{(3)}}{k^2}, \quad k \in \mathbb{N}, \tag{2.3}$$

   where $\{\beta_k^{(j)}\}_{-\infty, k\neq 0}^{\infty} \in l_2$ *for* $j = 1, 2, 3$, $B_0 = B_1 + B_2 + B_3$, *and* M_1 *and* M_2 *are the solutions (both real and not equal due to the inequalities* $B_j \neq B_p$ *for* $j \neq p$*) of equation* (1.50).

3. *The sequences* $\{\tau_k\}_{-\infty,k\ne0}^{\infty}$ *and* $\{\theta_k\}_{-\infty}^{\infty}\stackrel{\rm def}{=}\{0\}\bigcup\{\nu_k^{(1)}\}_{-\infty,k\ne0}^{\infty}\bigcup\{\nu_k^{(2)}\}_{-\infty,k\ne0}^{\infty}\bigcup\{\nu_k^{(3)}\}_{-\infty,k\ne0}^{\infty}$ *($\theta_{-k}=-\theta_k$, $\theta_k<\theta_{k+1}$) interlace in the following strict sense:*

$$\cdots<\theta_{-1}<\tau_{-1}<\theta_0=0<\tau_1<\theta_1<\tau_2<\cdots. \tag{2.4}$$

Then there exists a unique triplet $\{q_1(x),q_2(x),q_3(x)\}\in Q$ *such that the sequence* $\{\tau_k\}_{-\infty,k\ne0}^{\infty}$ *coincides with the spectrum of problem* (1.1j), (1.2j), (1.3), (1.4) *and the sequences* $\{\nu_k^{(j)}\}_{-\infty,k\ne0}^{\infty}$ ($j=1,2,3$) *coincide with the spectra of problems* (1.1j), (1.43j) *for* $j=1,2,3$, *respectively.*

Proof. Let us construct the following entire functions:

$$\tilde s_j(\lambda)=:\prod_1^{\infty}\left(1-\frac{\lambda^2}{\nu_k^{(j)2}}\right), \tag{2.5}$$

$$\phi_j(\lambda)=:\prod_1^{\infty}\left(1-\frac{\lambda^2}{\rho_k^{(j)2}}\right). \tag{2.6}$$

Due to [25, Lemma 2.1],

$$\tilde s_j(\lambda)=\frac{\sin\lambda a}{\lambda}-\frac{\pi B_j\cos\lambda a}{\lambda^2}+C^{(j)}\frac{\sin\lambda a}{\lambda^3}+\frac{f_j(\lambda)}{\lambda^3}, \tag{2.7}$$

$$\phi_1(\lambda)=\frac{\sin\lambda a}{\lambda}-\frac{\pi M_1\cos\lambda a}{\lambda^2}+E^{(1)}\frac{\sin\lambda a}{\lambda^3}+\frac{g_1(\lambda)}{\lambda^3}, \tag{2.8}$$

$$\phi_2(\lambda)=\frac{\sin\lambda a}{\lambda}-\frac{\pi M_2\cos\lambda a}{\lambda^2}+E^{(2)}\frac{\sin\lambda a}{\lambda^3}+\frac{g_2(\lambda)}{\lambda^3}, \tag{2.9}$$

where $C^{(j)}\in\mathbb{R}$, $f_j(\lambda)\in L_a$ for $j=1,2,3$, $E^{(j)}\in\mathbb{R}$, $g_j(\lambda)\in L_a$ for $j=1,2$. The following representation can be proved in the same way as Lemma 2.1 of [25]:

$$\phi_3(\lambda)=\cos\lambda a+\frac{\pi B_0\sin\lambda a}{3\lambda}+E^{(3)}\frac{\cos\lambda a}{\lambda^2}+\frac{g_3(\lambda)}{\lambda^2}, \tag{2.10}$$

where $E^{(3)}\in\mathbb{R}$ and $g_3(\lambda)\in L_a$.

Substituting (2.1) into (2.8)–(2.10), we obtain

$$\phi_1(\nu_k^{(1)})=(-1)^k\frac{a^2(B_1-M_1)}{\pi k^2}+\frac{\delta_k^{(4)}}{k^3}, \tag{2.11}$$

$$\phi_2(\nu_k^{(1)})=(-1)^k\frac{a^2(B_1-M_2)}{\pi k^2}+\frac{\delta_k^{(5)}}{k^3}, \tag{2.12}$$

$$\phi_3(\nu_k^{(1)})=(-1)^k\left(1-\frac{a^2B_1^2}{2k^2}+\frac{a^2B_0B_1}{3k^2}\right)+\frac{\delta_k^{(6)}}{k^3}, \tag{2.13}$$

$$\tilde s_2(\nu_k^{(1)})=(-1)^k\frac{a^2(B_1-B_2)}{\pi k^2}+\frac{\delta_k^{(7)}}{k^3}, \tag{2.14}$$

$$\tilde s_3(\nu_k^{(1)})=(-1)^k\frac{a^2(B_1-B_3)}{\pi k^2}+\frac{\delta_k^{(8)}}{k^3}, \tag{2.15}$$

where $\{\delta_k^{(j)}\}_{k=-\infty,k\ne0}^{\infty}\in l_2$ for $j=\overline{4,8}$.

Let us set

$$X_k^{(1)}=:\nu_k^{(1)}\left(\frac{3\phi_1(\nu_k^{(1)})\phi_2(\nu_k^{(1)})\phi_3(\nu_k^{(1)})}{\tilde s_2(\nu_k^{(1)})\tilde s_3(\nu_k^{(1)})}-\cos\nu_k^{(1)}a-\frac{\pi B_1\sin\nu_k^{(1)}a}{\nu_k^{(1)}}\right). \tag{2.16}$$

It is clear that $X_{-k} = -X_k$. To continue the proof we need the following lemma.

LEMMA 2.2.

$$\{X_k\}_{-\infty, k\neq 0}^{\infty} \in l_2. \tag{2.17}$$

Proof. To prove this lemma, it is enough to substitute (2.11)–(2.15) and the evident equality,

$$\cos \nu_k^{(1)} a + \frac{\pi B_1 \sin \nu_k^{(1)} a}{\nu_k^{(1)}} \underset{k\to\infty}{=} (-1)^k + O(k^{-2}),$$

into (2.16) and to take into account that M_1 and M_2 are the roots of (1.50). □

Now it follows from [19] that the series

$$\tilde{s}_1(\lambda) \sum_{\substack{k\neq 0 \\ -\infty}}^{\infty} \frac{X_k}{\frac{d\tilde{s}_1(\lambda)}{d\lambda}\Big|_{\lambda=\nu_k^{(1)}} (\lambda - \nu_k^{(1)})} \tag{2.18}$$

converges uniformly on any compact of the complex plane (and in the norm of $L_2(-\infty, \infty)$ for real λ) to a function $\varepsilon(\lambda)$, which belongs to L_a.

Introduce the function

$$R_1(\lambda) \stackrel{\text{def}}{=} \cos \lambda a + \frac{\pi B_1 \sin \lambda a}{\lambda} + \frac{\varepsilon(\lambda)}{\lambda}. \tag{2.19}$$

Since $\epsilon(\nu_k^{(1)}) = X_k^{(1)}$, (2.16) implies that

$$R_1(\nu_k^{(1)}) = \frac{3\phi_1(\nu_k^{(1)})\phi_2(\nu_k^{(1)})\phi_3(\nu_k^{(1)})}{\tilde{s}_2(\nu_k^{(1)})\tilde{s}_3(\nu_k^{(1)})}. \tag{2.20}$$

Let us denote the set of zeroes of the function $R_1(\lambda)$ by $\{\mu_k\}_{-\infty, k\neq 0}^{\infty}$. We number these zeroes as usual: $\mu_{-k} = -\mu_k$ and $\mu_k \leq \mu_{k+1}$. It follows from (2.19) (see [22, Lemma 3.4.2]) that

$$\mu_k^{(1)} \underset{k\to\infty}{=} \frac{\pi(2k-1)}{2a} + \frac{B_1}{k} + \frac{\gamma_k}{k}, \tag{2.21}$$

where $\{\gamma_k\}_{-\infty, k\neq 0}^{\infty} \in l_2$.

PROPOSITION 2.3.

$$\frac{3\phi_1(0)\phi_2(0)\phi_3(0)}{\tilde{s}_2(0)\tilde{s}_3(0)} > 0, \qquad (-1)^k \frac{3\phi_1(\nu_k^{(1)})\phi_2(\nu_k^{(1)})\phi_3(\nu_k^{(1)})}{\tilde{s}_2(\nu_k^{(1)})\tilde{s}_3(\nu_k^{(1)})} > 0. \tag{2.22}$$

Proof. The following inequalities can be obtained from (2.7)–(2.10) taking into account that all zeroes of $\phi_j(\lambda)$ (j=1,2,3) and all zeroes of $\tilde{s}_j(\lambda)$ are real:

$$\phi_j(0) > 0, \qquad \tilde{s}_j(0) > 0. \tag{2.23}$$

Hence the first of inequalities (2.22) follows. Let $\nu_k^{(1)} = \theta_p$, $(p \geq k > 0)$. Then it is clear from (2.4) and from the first of inequalities (2.22) that the function

$$\frac{3\phi_1(\lambda)\phi_2(\lambda)\phi_3(\lambda)}{\tilde{s}_1(\lambda)\tilde{s}_2(\lambda)\tilde{s}_3(\lambda)}$$

is positive (negative) on intervals (θ_k, τ_{k+1}) (on intervals (τ_k, θ_k)). Hence

$$\frac{3\phi_1(\theta_p)\phi_2(\theta_p)\phi_3(\theta_p)}{\tilde{s}_2(\theta_p)\tilde{s}_3(\theta_p)} \lim_{\lambda\to\theta_p} \frac{(\lambda-\theta_p)}{\tilde{s}_1(\lambda)} = \lim_{\lambda\to\theta_p} \frac{3\phi_1(\lambda)\phi_2(\lambda)\phi_3(\lambda)(\lambda-\theta_p)}{\tilde{s}_2(\lambda)\tilde{s}_3(\lambda)\tilde{s}_1(\lambda)} > 0. \tag{2.24}$$

From the other hand,

$$\lim_{\lambda\to\theta_p} \frac{(\lambda-\theta_p)}{\tilde{s}_1(\lambda)}(-1)^k = \lim_{\lambda\to\nu_k^{(1)}} \frac{(\lambda-\nu_k^{(1)})}{\tilde{s}_1(\lambda)}(-1)^k > 0. \tag{2.25}$$

The second of inequalities (2.22) follows from (2.24) and (2.25). □

Using (2.19), (2.20), and (2.22), we obtain

$$R_1(0) > 0, \quad (-1)^k R_1(\nu_k^{(1)}) > 0. \tag{2.26}$$

These inequalities imply the following proposition.

PROPOSITION 2.4.

$$\cdots < \nu_{-1}^{(1)} < \mu_{-1}^{(1)} < 0 < \mu_1^{(1)} < \nu_1^{(1)} < \mu_2^{(1)} < \nu_2^{(1)} < \cdots . \tag{2.27}$$

Now the two sequences $\{\nu_k^{(1)}\}_{-\infty,k\neq0}^{\infty}$ *and* $\{\mu_k^{(1)}t\}_{-\infty,k\neq0}^{\infty}$ *satisfy (due to* (2.1), (2.21)*, and Proposition* 2.4*) the conditions of* [22, *Theorem* 3.4.1]. *Thus it is possible to construct (via the well-known procedure* [21], [22, *section* 3.4]*) a unique real* $q_1(x) \in L_2(0,a)$ *such that* $\{\nu_k^{(1)}\}_{-\infty,k\neq0}^{\infty}$ *is the spectrum of problem* (1.1_1), (1.13) *and* $\{\mu_k^{(1)}\}_{-\infty,k\neq0}^{\infty}$ *is the spectrum of the problem*

$$\begin{aligned} y_1'' + \lambda^2 y_1 - q_1(x)y_1 = 0, \\ y_1(\lambda,0) = y_1'(\lambda,a) = 0. \end{aligned} \tag{2.28}$$

In the same way we can construct $q_2(x)$ *and* $q_3(x)$ *(each of them is unique real and belongs to* $L_2(0,a)$*). It is clear that the obtained triplet* $\{q_1(x), q_2(x), q_3(x)\}$ *generates the spectra of problems* $(1.1j)$, $(1.43j)$*, which coincide with* $\{\nu_k^{(j)}\}_{-\infty,k\neq0}^{\infty}$ *for* $j = 1,2,3$*, respectively, and the functions* $s_j(\lambda,a)$ *which coincide with* $\tilde{s}_j(\lambda)$ *defined by* (2.5)*. The sets of zeroes of values of the derivatives* $s_j'(\lambda,a)$ *coincide with* $\{\mu_k^{(j)}\}_{-\infty,k\neq0}^{\infty}$*, and consequently* $s_j'(\lambda,a)$ *coincide with* $R_j(\lambda)$*, where* $R_1(\lambda)$ *is defined by* (2.19)*. (The expressions for* $R_2(\lambda)$ *and* $R_3(\lambda)$ *are analogous.) Thus the values of the function*

$$\tilde{\varphi}_1(\lambda) \stackrel{\text{def}}{=} s_1(\lambda,a)s_2(\lambda,a)s_3'(\lambda,a) + s_2(\lambda,a)s_3(\lambda,a)s_1'(\lambda,a) + s_1(\lambda,a)s_3(\lambda,a)s_2'(\lambda,a)$$

at $\lambda = \nu_k^{(j)}$ *coincide with* $3\phi_1(\nu_k^{(j)})\phi_2(\nu_k^{(j)})\phi_3(\nu_k^{(j)})$ *for all* $k = \pm1, \pm2, \ldots$ *and all* $j = 1,2,3$*, i.e., with the corresponding values of the function* $3\phi_1(\lambda)\phi_2(\lambda)\phi_3(\lambda)$*. This means that the spectrum of problem* $(1.1j)$, $(1.2j)$, (1.3), (1.4) *generated by obtained triplet* $\{q_1(x), q_2(x), q_3(x)\}$ *coincides with* $\{\tau_k\}_{-\infty,k\neq0}^{\infty}$.

The uniqueness of the solution of the inverse problem follows from the fact that formula (2.18) *establishes one-to-one correspondence between* l_2 *and* L_a *(see* [19]).

Remark 2.1. If the spectra intersect, i.e., condition 1(ii) of Theorem 2.1 is violated (and consequently condition 3 is, too), then the solution of the inverse problem is not unique because of the same reasons as in the case of three spectral problem (see [11], [25]).

Let us now consider Hochstadt–Lieberman type (see [15], [12], [6]) inverse problem, i.e., inverse problem with partial information on the potential.

THEOREM 2.5. *The two potentials $q_1(x)$ and $q_2(x)$ (both real-valued and belonging to $L_2(0,a)$) and the spectrum $\{\tau_k\}_{-\infty,k\neq 0}^{\infty}$ of the problem (1.1j), (1.2j) ($j=1,2,3$), (1.3), (1.4) determine uniquely the (real-valued and belonging to $L_2(0,a)$) potential $q_3(x)$.*

Proof. Let us suppose there exist two potentials $q_3(x)$ and $\tilde{q}_3(x)$ (both real-valued and $\in L_2(0,a)$) such that problems (1.1j), (1.2j), (j=1,2,3), (1.3), (1.4) generated by the triplets $\{q_1(x),q_2(x),q_3(x)\}$ and $\{q_1(x),q_2(x),\tilde{q}_3(x)\}$ possess the same spectrum $\{\tau_k\}_{-\infty,k\neq 0}^{\infty}$. Then using (1.7)–(1.9), we obtain

$$\varphi_1(\lambda)=s_1(\lambda,a)s_2(\lambda,a)s_3'(\lambda,a)+s_1(\lambda,a)s_3(\lambda,a)s_2'(\lambda,a)+s_2(\lambda,a)s_3(\lambda,a)s_1'(\lambda,a), \tag{2.29}$$

$$\varphi_1(\lambda)=s_1(\lambda,a)s_2(\lambda,a)\tilde{s}_3'(\lambda,a)+s_1(\lambda,a)\tilde{s}_3(\lambda,a)s_2'(\lambda,a)+s_2(\lambda,a)\tilde{s}_3(\lambda,a)s_1'(\lambda,a), \tag{2.30}$$

where $\varphi_1(\lambda)$ is determined uniquely by the spectrum $\{\tau_k\}_{-\infty,k\neq 0}^{\infty}$ via the formulas $\varphi_1(\lambda)=3\phi_1(\lambda)\phi_2(\lambda)\phi_3(\lambda)$ and (2.6). The functions $s_1(\lambda,a)$, $s_2(\lambda,a)$, $s_1'(\lambda,a)$ and $s_2'(\lambda,a)$ are determined by $q_1(x)$ and $q_2(x)$. (They can be found by solving the corresponding Dirichlet and Dirichlet–Neumann problems.) Now we can find the values $s_3(\nu_k^{(1)},a)$ and $\tilde{s}_3(\nu_k^{(1)},a)$, where $\nu_k^{(1)}$ are the zeroes of $s_1(\lambda,a)$ via the formula

$$\tilde{s}_3(\nu_k^{(1)},a)=s_3(\nu_k^{(1)},a)=\frac{\varphi_1(\nu_k^{(1)})}{s_2(\nu_k^{(1)},a)s_1'(\nu_k^{(1)},a)} \tag{2.31}$$

if $s_2(\nu_k^{(1)},a)\neq 0$. In the opposite case (if $s_2(\nu_k^{(1)},a)=0$), formula (2.31) should be replaced by the following:

$$\tilde{s}_3(\nu_k^{(1)},a)=s_3(\nu_k^{(1)},a)=\lim_{\lambda\to\nu_k^{(1)}}\frac{\varphi_1(\lambda)}{s_2(\lambda,a)s_1'(\lambda,a)+s_1(\lambda,a)s_2'(\lambda,a)}. \tag{2.32}$$

The limit exists because: (1) all zeroes of $s_j(\lambda,a)$ are simple; (2) if $s_j(\lambda,a)=0$, then $s_j'(\lambda,a)\neq 0$; (3) according to Theorem 1.13, the equation $s_1(\lambda,a)=s_2(\lambda,a)=0$ implies $\varphi_1(\lambda)=0$; (4) all zeroes of the function $\psi_1(\lambda)=s_2(\lambda,a)s_1'(\lambda,a)+s_1(\lambda,a)s_2'(\lambda,a)$ are simple due to Lemma 1.1; and (5) if $\psi_1(\nu_k^{(1)})=0$, then due to Lemma 1.15, $\varphi_1(\nu_k^{(1)})=0$. Equations (2.29) and (2.30) imply that

$$\begin{aligned}&\nu_k^{(1)2}\left(\tilde{s}_3(\nu_k^{(1)},a)-\frac{\sin\nu_k^{(1)}a}{\nu_k^{(1)}}+\frac{\pi B_1\cos\nu_k^{(1)}a}{\nu_k^{(1)2}}\right)\\ &\quad=\nu_k^{(1)2}\left(s_3(\nu_k^{(1)},a)-\frac{\sin\nu_k^{(1)}a}{\nu_k^{(1)}}+\frac{\pi B_1\cos\nu_k^{(1)}a}{\nu_k^{(1)2}}\right)\overset{\text{def}}{=}d_k\end{aligned} \tag{2.33}$$

for all k. Using (1.17) and [22, Lemma 1.4.3], we obtain $\{d_k\}_{-\infty,k\neq 0}^{\infty}\in l_2$. This means that

$$\begin{aligned}\lambda^2\left(\tilde{s}_3(\lambda,a)-\frac{\sin\lambda a}{\lambda}+\frac{\pi B_1\cos\lambda a}{\lambda^2}\right)&=\lambda^2\left(s_3(\lambda,a)-\frac{\sin\lambda a}{\lambda}+\frac{\pi B_1\cos\lambda a}{\lambda^2}\right)\\ &=s_1(\lambda,a)\sum_{\substack{k\neq 0\\ -\infty}}^{\infty}\frac{d_k}{\frac{ds_1(\lambda,a)}{d\lambda}\Big|_{\lambda=\nu_k^{(1)}}(\lambda-\nu_k^{(1)})},\end{aligned} \tag{2.34}$$

and consequently $s_3(\lambda,a)=\tilde{s}_3(\lambda,a)$. Substituting this equation into (2.29) and (2.30), we obtain the identity $s_3'(\lambda,a)=\tilde{s}_3'(\lambda,a)$. The sets of zeroes of the functions $s_3(\lambda,a)$ and $s_3'(\lambda,a)$ uniquely determine $q_3(x)$ ([3], [21], [22]). □

The following theorem can be proved in the same way.

THEOREM 2.6. *Let the following data be given:* (1) *a (real-valued belonging to* $L_2(0,a)$*) potential* $q_1(x)$, (2) *the spectrum* $\{\tau_k\}_{-\infty,k\neq0}^{\infty}$ *of problem* (1.1j), (1.2j), (1.3), (1.4), *and* (3) *the spectrum* $\{\nu_k^{(2)}\}_{-\infty,k\neq0}^{\infty}$ *of problem* (1.1$_2$), (1.14).

If $\{\nu_k^{(2)}\}_{-\infty,k\neq0}^{\infty}\bigcap\{\nu_k^{(1)}\}_{-\infty,k\neq0}^{\infty}=\emptyset$, $\{\nu_k^{(2)}\}_{-\infty,k\neq0}^{\infty}\bigcap\{\tau_k\}_{-\infty,k\neq0}^{\infty}=\emptyset$, *and* $\{\nu_k^{(1)}\}_{-\infty,k\neq0}^{\infty}\bigcap\{\tau_k\}_{-\infty,k\neq0}^{\infty}=\emptyset$, *then these data determine uniquely the potentials* $q_2(x)$ *and* $q_3(x)$ (*in the class of real-valued* $L_2(0,a)$ *functions*).

Proof. Since given $q_1(x)$ uniquely determine the functions $s_1(\lambda,a)$ and $s_1'(\lambda,a)$ and the spectrum $\{\nu_k^{(2)}\}_{-\infty,k\neq0}^{\infty}$ uniquely determine the function $s_2(\lambda,a)$, formula (2.32) uniquely determine the set $\{s_3(\nu_k^{(1)},a)\}_{-\infty,k\neq0}^{\infty}$. Thus formula (2.33) uniquely determine the set $\{d_k\}_{-\infty,k\neq0}^{\infty}$. Consequently, formula (2.34) uniquely determine $s_3(\lambda,a)$ and the set of its zeroes $\{\nu_k^{(3)}\}_{-\infty,k\neq0}^{\infty}$. Now it is sufficient to apply Theorem 2.1 to finish the proof. □

REFERENCES

[1] V. ADAMYAN, *Scattering Matrices for Microschemes*, Oper. Theory Adv. Appl. 59, Birkhäuser, Basel, 1992.

[2] V.M. ADAMYAN AND B.P. PAVLOV, *Null-range potentials and M.G. Krein's formula of generalized resolvents*, in Studies on Linear Operators of Functions XV, Research Notes of Scientific Seminar of the Leningrad Branch of Mathematical Institute 149, the Steklov Institute, St. Petersburg, 1986, pp. 7–23 (in Russian).

[3] G. BORG, *Uniqueness theorems in the spectral theory of* $y''+(\lambda-q(x))y=0$, in Proceedings of the 11th Scandinavian Congress of Mathematicians, Johan Grundt Tanums Forlag, Oslo, 1952, pp. 276–287.

[4] V.B. BOGEVOLNOV, A.B. MIKHAILOVA, B.S. PAVLOV, AND A.M. YAFYASOV, *About Scattering on the Ring*, Oper. Theory Adv. Appl., Birkhäuser, Basel, Boston, to appear.

[5] R. COURNAT AND D. HILBERT, *Methods of Mathematical Physics* 1, Interscience, New York, 1953.

[6] R. DEL RIO, F. GESZTESY, AND B. SIMON, *Inverse spectral analysis with partial information on the potential* III: *Updating boundary conditions*, Internat. Math. Res. Notices, 15 (1997), pp. 751–758.

[7] P. EXNER AND P. SEBA, *A new type of quantum interference transistor*, Phys. Lett. A, 129 (1988), pp. 477–480.

[8] P. EXNER AND R. GAWLISTA, *Band spectra of rectangular graph superlattices*, Phys. Rev. B, 53 (1996), pp. 4275–4286.

[9] I.C. GOHBERG AND M.G. KREIN, *Introduction to the Theory of Linear Nonselfadjoint Operators in Hilbert Space*, AMS, Providence, RI, 1969.

[10] I.C. GOHBERG, S. GOLDBERG, AND M.A. KAASHOEK, *Classes of Linear Operators* I, Oper. Theory Adv. Appl. 49, Birkhäuser, Basel, 1990.

[11] F. GESZTESY AND B. SIMON, *On the determination of a potential from three spectra*, in Advances in Mathematical Sciences, V. Buslaev and M. Solomyak, eds., Amer. Math. Soc. Transl. Ser. 2 189, AMS, Providence, RI, 1999, pp. 85–92.

[12] F. GESZTESY AND B. SIMON, *Inverse spectral analysis with partial information on the potential* II: *The case of discrete spectrum*, Trans. Amer. Math. Soc., 352 (1999), pp. 2765–2787.

[13] N.I. GERASIMENKO, *Inverse scattering problem on noncompact graph*, Teoret. Mat. Fiz., 75 (1988), pp. 187–200 (in Russian).

[14] N.I. GERASIMENKO AND B.S. PAVLOV, *Scattering problem on noncompact graphs*, Teoret. Mat. Fiz., 74 (1988), pp. 345–359 (in Russian).

[15] H. HOCHSTADT AND B. LIEBERMAN, *An inverse Sturm-Liouville problem with mixed given data*, SIAM J. Appl. Math., 34 (1978), pp. 676–680.

[16] V. Kostrykin and R. Schrader, *Kirchhoff's rule for quantum wires*, J. Phys. A, 32 (1999), pp. 595–630.
[17] M.G. Krein and H. Langer, *On some mathematical principles in the linear theory of damped oscillations of continua* II, Integral Equations Operator Theory, 1 (1978), pp. 539–566.
[18] B.Ja. Levin, *Lectures on Entire Functions*, Transl. Math. Monogr. 150, AMS, Providence, RI, 1996.
[19] B.Ja. Levin and Yu. I. Lyubarskii, *Interpolation by entire functions of special classes and related expansions in series of exponents*, Izv. Acad. Sci. USSR Ser. Mat., 39 (1975), pp. 657–702 (in Russian).
[20] B.Ja. Levin and I.V. Ostrovskii, *On small perturbations of the set of roots of a sinus-type function*, Izv. Akad. Nauk USSR Ser. Mat., 43 (1979), pp. 87–110 (in Russian).
[21] B.M. Levitan and M.G. Gasymov, *Determination of differential equation by two spectra*, Uspechi Math. Nauk, 19 (1964), pp. 3–63 (in Russian).
[22] V.A. Marchenko, *Sturm-Liouville Operators and Applications*, Oper. Theory Adv. Appl. 22, Birkhäuser, Basel, Boston, 1986.
[23] V. Pivovarchik, *Scattering in a Loop-Shaped Waveguide*, Oper. Theory Adv. Appl., Birkhäuser, Basel, Boston, to appear.
[24] V. Pivovarchik, *Inverse problem on a semiinfinite graph*, Funct. Anal. Appl., submitted (in Russian).
[25] V. Pivovarchik, *An inverse Sturm-Liouville problem by three spectra*, Integral Equations Operator Theory, 34 (1999), pp. 234–243.
[26] V. Pivovarchik, *Direct and inverse three-point Sturm-Liouville problem with parameter-dependent boundary conditions*, Asymptot. Anal., to appear.

SIAM J. MATH. ANAL.
Vol. 32, No. 4, pp. 820–853

UNFOLDING SINGULARLY PERTURBED BOGDANOV POINTS*

MATTHIAS STIEFENHOFER†

Abstract. Bogdanov points that occur in the fast dynamics of singular perturbation problems are often encountered in applications; e.g., in the van der Pol–Duffing oscillator [M. Koper, *Phys. D*, 80 (1995), pp. 64–88] or in the FitzHugh–Nagumo equation [W.-J. Beyn and M. Stiefenhofer, *J. Dynam. Differential Equations*, 11 (1997), pp. 671–709]. A parameter of the normal form is taken to be a dynamic variable with slow dynamics of speed ϵ. We analyze these points using a generic unfolding which ensures that the typical phenomena near a regularly perturbed Bogdanov point (saddle-nodes, Hopf points, periodic orbits, homoclinic orbits) carry over to Bogdanov points viewed in the context of singular perturbations. We combine analytical and numerical results to study the relations between these structures in the 3-dimensional unfolding space. In particular, the singularly perturbed homoclinic orbits can be analyzed after an appropriate blow-up of the singularly perturbed Bogdanov point using a technique from [W.-J. Beyn and M. Stiefenhofer, *J. Dynam. Differential Equations*, 11 (1997), pp. 671–709]. As indicated by numerical calculations the homoclinic orbits turn into homoclinic orbits of Shilnikov type that vanish presumably by a canard like explosion process [M. Diener, *Étude générique des canards*, IRMA, 1981], [W. Eckhaus, Lecture Notes in Math. 985, Springer, 1983, pp. 449–494], [V. I. Arnold, V. S. Afrajmovich, Y. S. Il'yashenko, and L. P. Shil'nikov, *Dynamical Systems* V: *Bifurcation Theory*, Springer, 1994], [F. Dumortier and R. Roussarie, Mem. Amer. Math. Soc. 121, AMS, 1996].

Key words. Bogdanov points, homoclinic orbits, singular perturbations

AMS subject classifications. 34C37, 34E15, 58F14

PII. S0036141098334237

1. Introduction.

Singular perturbation problems of the form

$$\begin{aligned} \dot{u} &= F(u,v) \in \mathbb{R}^n, \\ \dot{v} &= \epsilon \cdot G(u,v) \in \mathbb{R}, \qquad -1 \ll \epsilon \ll 1 \end{aligned} \tag{1.1}$$

represent one possibility of including the real parameter v of the u-system $\dot{u} = F(u,v)$ into the dynamics of the model. We investigate this perturbation process from $\epsilon = 0$ to $\epsilon \neq 0$ when the central u-system $\dot{u} = F(u,0)$ has a Bogdanov point, i.e., a stationary point with an algebraically double and geometrically simple eigenvalue zero [15]. We refer to a point of this kind as a singularly perturbed Bogdanov point of the fast-slow system (1.1) if it additionally satisfies $G(u,0) = 0$. In this situation the well-known technique of reducing to the 1-dimensional slow manifold defined by $F(u,v) = 0$ is impossible (cf. [12], [29]) and instead of enslaving the fast u-variables by the slow v-variable (cf. [17]) we may obtain new phenomena due to the equal coupling of all variables.

The standard examples of singularly perturbed Bogdanov points are given by the van der Pol–Duffing oscillator [19]

$$\begin{aligned} \dot{u}_1 &= p_2 u_2 - u_1^3 + 3u_1 - p_1, \\ \dot{u}_2 &= u_1 - 2u_2 + v, \\ \dot{v} &= \epsilon \cdot (u_2 - v), \end{aligned} \tag{1.2}$$

*Received by the editors February 17, 1998; accepted for publication (in revised form) July 24, 2000; published electronically December 13, 2000. This work was supported by DFG Schwerpunktprogramm "Ergodentheorie, Analysis und effiziente Simulation dynamischer Systeme."
http://www.siam.org/journals/sima/32-4/33423.html
†Fakultät für Mathematik, Universität Bielefeld, Postfach 100131, 33501 Bielefeld, Germany. Current address: Martinstrasse 19, 88161 Lindenberg/Allgäu, Germany.

and the travelling wave problem for the FitzHugh–Nagumo equation [5], [16], [21]

$$\begin{aligned} \dot{u}_1 &= u_2, \\ \dot{u}_2 &= cu_2 + u_1(u_1 - a)(u_1 - 1) + v, \\ \dot{v} &= \frac{\bar{\epsilon}}{c} \cdot (u_1 - \gamma v). \end{aligned} \tag{1.3}$$

The van der Pol–Duffing oscillator has two singularly perturbed Bogdanov points

$$(u_1, u_2, v, p_1, p_2, \epsilon) = \left(\pm\frac{1}{\sqrt{3}}, \pm\frac{1}{\sqrt{3}}, \pm\frac{1}{\sqrt{3}}, \mp\frac{4}{3\sqrt{3}}, -4, 0 \right), \tag{1.4}$$

whereas the travelling wave problem (1.3) (with $\epsilon = \bar{\epsilon}/c$) has one singularly perturbed Bogdanov point at $(u, v, a, c, \epsilon) = 0$. In [4] we examined the travelling wave problem (1.3) with respect to homoclinic orbits near the singularly perturbed Bogdanov point.

In the current paper we perform two extensions. First, we extend the results from [4] to general systems of the form (1.1). For this purpose the notion of a generic unfolding of a singularly perturbed Bogdanov point is introduced. In that framework we do not restrict ourselves to homoclinic orbits but investigate to a certain extent the unfolding structure of stationary points, periodic orbits, homoclinic orbits, and invariant tori. We use analytical results from [11], [10], [4], [26], [27], numerical results from [19], and heuristic results from [25] that are mainly obtained from the model equation

$$\begin{aligned} \dot{u}_1 &= u_2, \\ \dot{u}_2 &= u_1^2 + u_1 u_2 + k u_2 + v, \\ \dot{v} &= \epsilon \cdot (\lambda - u_1) \end{aligned} \tag{1.5}$$

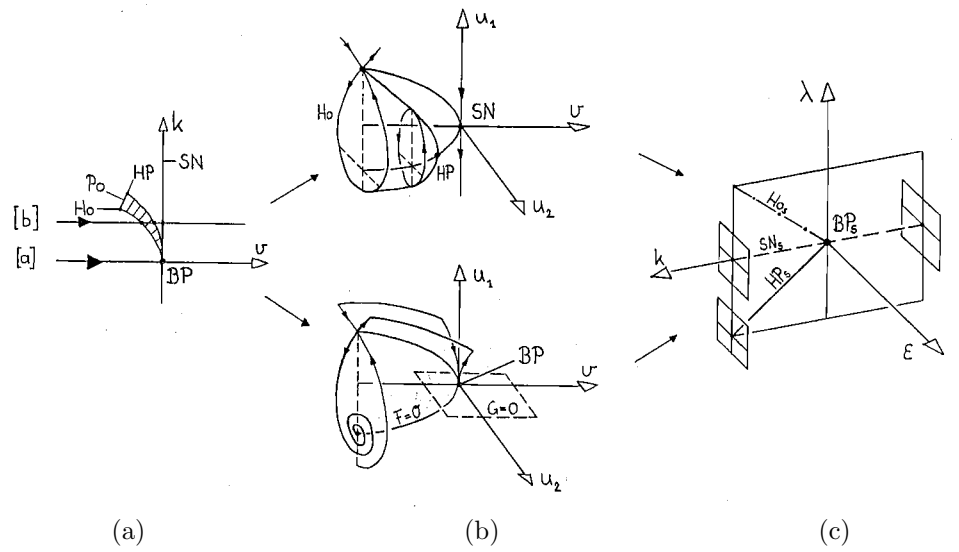
which represents a generic (λ, k)-unfolding of the singularly perturbed Bogdanov point $(u, v, \lambda, k, \epsilon) = 0$. Other examples of singularly perturbed Bogdanov points appear in the context of cell communication problems [14], [25].

The model equation (1.5) indicates our aim; for $\epsilon = 0$ we have a family of fast u-systems $\dot{u} = F(u, v, k)$ that depends on the two parameters v and k. The well-known (v, k)-unfolding diagram with corresponding bifurcation diagrams along sections [a], [b] is shown in the left and middle diagram of Figure 1 (cf. [15]).

The parameters v and k are of different nature in the sense that the v-axis in diagram (a) transversally intersects the curve of saddle-nodes SN; i.e., the v-axis represents a generic parameter direction in the 2-dimensional (v, k)-unfolding space of the Bogdanov point BP. Note also that diagram (a) shows the universal unfolding of a finitely determined Bogdanov point that occurs in the fast u-system; therefore Figure 1 shows the standard case [15]. The fast u-system of the van der Pol–Duffing oscillator (1.2) corresponds to the standard case.

By the equation $\dot{v} = \epsilon(\lambda - u_1) = \epsilon G(u, \lambda)$ the generic parameter direction v is slowly included into the dynamics of the system (1.5). Graphically, we perturb the fast dynamics shown in diagram (b) of Figure 1 by a slow dynamics in the direction of v.

This perturbation is not done in an arbitrary way; it is controlled with the help of the additional parameter λ which ensures that at $\lambda = 0$ the nullclines of the u- and v-equations meet transversally in the Bogdanov point. Compare $F = 0$ and $G = 0$ in the lower half of diagram (b) in Figure 1.



(a) (b) (c)

FIG. 1. (a) *The (v,k)-unfolding space of the fast u-system, BP = Bogdanov point, SN = saddle-node, HP = Hopf point, Po = periodic orbit, Ho = homoclinic orbit;* (b) *the bifurcation diagrams along sections* [a] *and* [b]*;* (c) *the* 3*-dimensional (λ,k,ϵ)-unfolding space of the fast-slow system* (1.5) *(BP_s, SN_s, HP_s, Ho_s explained in text).*

Note also that the curve $F=0$ can either be interpreted as a branch of stationary points of the fast u-system $\dot{u}=F(u,v,k)$ or as a continuum of stationary points in the (u,v)-phase space of the fast-slow system $\dot{u}=F(u,v,k)$, $\dot{v}=0$. The continuum of stationary points represents the singular character of our perturbation process.

With these assumptions we try to understand the structures in the 3-dimensional (λ,k,ϵ)-unfolding space of a singularly perturbed Bogdanov point (cf. (c) in Figure 1).

The paper is organized as follows. In section 2 the definitions of singularly perturbed saddle-nodes SN_s, singularly perturbed Hopf points HP_s, singularly perturbed homoclinic orbits Ho_s, and singularly perturbed Bogdanov points BP_s are introduced together with the notion of a generic unfolding. The relations between these points are summarized in a theorem and indicated in the (λ,k)-plane of Figure 1(c).

At this stage the first differences between the standard case and the FitzHugh–Nagumo equation (1.3) occur. The Bogdanov point $(u,v,a,c)=0$ of the fast u-system in (1.3) is not finitely determined; i.e., one of the two normal form coefficients vanishes which results from the fact that the central u-system $\dot{u}_1=u_2$, $\dot{u}_2=-u_1^2+u_1^3$ of the FitzHugh–Nagumo equation is Hamiltonian. The theorem applies for the standard case and the FitzHugh–Nagumo equation as well.

Next we perform the actual singular perturbation; i.e., we perturb the points SN_s, HP_s, Ho_s, and BP_s from the (λ,k)-plane in Figure 1(c) to $\epsilon\neq 0$. Here we combine analytical, numerical, and heuristical results for deriving a rough version of the complete (λ,k,ϵ)-unfolding diagram. The underlying analytical results from [8], [11], [1], [10], [26], [27] and [4] are summarized in three theorems in sections 4 and 5.

More precisely, in section 4 the unfolding structures of singularly perturbed saddle-nodes SN_s and singularly perturbed Hopf points HP_s are investigated. These unfolding diagrams represent 2-dimensional sections through the 3-dimensional unfolding space of a singularly perturbed Bogdanov point BP_s. Some of the sections

are indicated in Figure 1(c). Section 5 deals with singularly perturbed homoclinic orbits Ho_s obtained by blowing up the singularly perturbed Bogdanov point BP_s appropriately; i.e., here we do not restrict ourselves to 2-dimensional sections, but we start from the center of the (λ, k, ϵ)-unfolding diagram. Section 5 contains the most technical part of the paper.

We conclude the paper with some strange results and conjectures concerning a canard like behavior [8], [11], [1], [10] of homoclinic orbits of Shilnikov type motivated by numerical calculations.

2. The setting and the unfolding diagram. Consider a system of the form

$$\begin{aligned} \dot{u} &= F(u, v, \epsilon) \in \mathbb{R}^2, \\ \dot{v} &= \epsilon \cdot G(u, v, \epsilon) \in \mathbb{R}. \end{aligned} \tag{2.1}$$

Concerning the fast u-system $\dot{u} = F(u, 0, 0)$ at $v = 0$ assume

$$F^0 = 0, \qquad \operatorname{tr}(F_u)^0 = 0, \qquad \det(F_u)^0 = 0, \qquad F_u^0 \neq 0 \in \mathbb{R}^{2,2}, \tag{2.2}$$

where tr, det denote the trace, determinant of a matrix and the upper index "0" denotes evaluation at $(u, v, \epsilon) = (0, 0, 0)$. Concerning the regular v-perturbation $\dot{u} = F(u, v, 0)$ assume

$$F_v^0 \notin \mathrm{R}[F_u^0], \qquad F_{uu}^0 \varphi_1^2 \notin \mathrm{R}[F_u^0] \qquad \text{with} \qquad \mathrm{N}[F_u^0] = \operatorname{span}\{\varphi_1\}, \tag{2.3}$$

where R and N denote the range and kernel of a matrix, respectively. Finally, concerning the singular v-perturbation $\dot{v} = \epsilon G(u, v, \epsilon)$ assume

$$G^0 = 0, \qquad \begin{pmatrix} F_u & F_v \\ G_u & G_v \end{pmatrix}^0 \quad \text{nonsingular .} \tag{2.4}$$

We interpret these assumptions. According to (2.2) the central u-system $\dot{u} = F(u, 0)$ has a Bogdanov point in the sense defined in the introduction. According to (2.3) the regular v-perturbation $\dot{u} = F(u, v, 0)$ of this system has a quadratic limit point with respect to v; i.e., from a stationary point of view the Bogdanov point is generically perturbed by the parameter v. In the next step, the parameter v of the u-system $\dot{u} = F(u, v, 0)$ is included into the dynamics of the (u, v)-system (2.1) according to $\dot{v} = \epsilon G(u, v, \epsilon)$ where this dynamical unfolding is characterized by (2.4); i.e., the F-nullcline defined by $F(u, v, 0) = 0 \in \mathbb{R}^2$ and the surface defined by $G(u, v, 0) = 0 \in \mathbb{R}$ intersect transversally in the Bogdanov point $(u, v) = 0$. Typically, we arrive at a configuration as depicted in the lower part of Figure 1(b).

Definition 2.1. *A point $(u, v, \epsilon) = 0$ which satisfies (2.2)–(2.4) is called a singularly perturbed Bogdanov point of the system (2.1).*

Summing up we investigate a certain dynamical unfolding of a Bogdanov point of the fast u-system which satisfies $G = 0$ and some additional nondegeneracy conditions.

Next we turn to the assumptions concerning the parameter unfolding of system (2.1). Apart from the basic condition $(F, G)^0 = 0$ the central point $(u, v, \epsilon) = 0$ satisfies the two degeneracy conditions $\operatorname{tr}(F_u)^0 = 0$ and $\det(F_u)^0 = 0$. Hence it seems appropriate to examine a singularly perturbed Bogdanov point within a parameter unfolding of two parameters $p = (p_1, p_2)$:

$$\begin{aligned} \dot{u} &= F(u, v, p_1, p_2, \epsilon) \in \mathbb{R}^2, \\ \dot{v} &= \epsilon \cdot G(u, v, p_1, p_2, \epsilon) \in \mathbb{R}. \end{aligned} \tag{2.5}$$

The corresponding nondegeneracy condition of this parameter unfolding is determined with the help of the defining equation

$$H(u,v,p_1,p_2) := \begin{pmatrix} F(u,v,p_1,p_2,0) \\ G(u,v,p_1,p_2,0) \\ \det(F_u)(u,v,p_1,p_2,0) \\ \operatorname{tr}(F_u)(u,v,p_1,p_2,0) \end{pmatrix} = 0, \qquad H:\ \mathbb{R}^5 \to \mathbb{R}^5. \tag{2.6}$$

DEFINITION 2.2. *The (p_1,p_2)-unfolding* (2.5) *of system* (2.1) *is a generic unfolding of a singularly perturbed Bogdanov point if $H'(0) \in \mathbb{R}^{5,5}$ is nonsingular.* In detail, the linearization $H'(0)$ is given by

$$H'(0) = \begin{pmatrix} F_u & F_v & F_{p_1} & F_{p_2} \\ G_u & G_v & G_{p_1} & G_{p_2} \\ \det_u(F_u) & \det_v(F_u) & \det_{p_1}(F_u) & \det_{p_2}(F_u) \\ \operatorname{tr}_u(F_u) & \operatorname{tr}_v(F_u) & \operatorname{tr}_{p_1}(F_u) & \operatorname{tr}_{p_2}(F_u) \end{pmatrix}^0 \in \mathbb{R}^{5,5}. \tag{2.7}$$

Hence, the point $(u,v,p_1,p_2) = 0$ must be a regular solution of the defining equation (2.6). This condition simply ensures that the singularly perturbed Bogdanov point $(u,v,\epsilon) = 0$ of system (2.1) is unique within the unfolded system (2.5).

The van der Pol–Duffing oscillator (1.2), the travelling wave problem (1.3) (with $(p_1,p_2) = (a,c)$), and the model equation (1.5) (with $(p_1,p_2) = (\lambda,k)$) satisfy the conditions (2.2)–(2.4) and (2.7) at the Bogdanov points. Further examples can be found in [25] where cell communication problems from [14] are studied.

Our aim is to determine the unfolding structure of the general system (2.5) in the (p_1,p_2,ϵ)-parameter space. For this purpose we first determine the unfolding structure of the u-system $\dot u = F(u,v,p_1,p_2,0)$ and then we perform the singular perturbation from $\epsilon = 0$ to $\epsilon \neq 0$; i.e., we include the generic parameter v into the dynamics of the system according to $\dot v = \epsilon G(u,v,p_1,p_2,\epsilon)$.

Now the regularity of $H'(0)$ ensures that the singularly perturbed Bogdanov point lies at the intersection of the curves defined by

$$[F,G,\det(F_u)](u,v,p_1,p_2,0) = 0 \quad \text{and} \quad [F,G,\operatorname{tr}(F_u)](u,v,p_1,p_2,0) = 0. \tag{2.8}$$

In section 4 it is shown that these curves give rise to saddle-node and Hopf points of the u-system $\dot u = F(u,v,p_1,p_2,0)$ which satisfy $G(u,v,p_1,p_2,0) = 0$. Moreover, we shall see that a curve of homoclinic base points (the steady states to which the homoclinic orbits converge) of the system $\dot u = F(u,v,p_1,p_2,0)$ with $G(u,v,p_1,p_2,0) = 0$ emanates from the singularly perturbed Bogdanov point.

Analogous to the definition of a singularly perturbed Bogdanov point, we refer to these phenomena as *singularly perturbed saddle-nodes, Hopf points, and homoclinic base points*, respectively. In this sense the assumptions (2.2)–(2.4) and (2.7) guarantee that the typical phenomena near a regularly perturbed Bogdanov point carry over to Bogdanov points viewed in the context of singular perturbations. The precise definitions of singularly perturbed saddle-node and Hopf points (which are of lower degeneracy than singularly perturbed Bogdanov points) are given in section 4. Singularly perturbed homoclinic orbits are precisely defined in section 5.

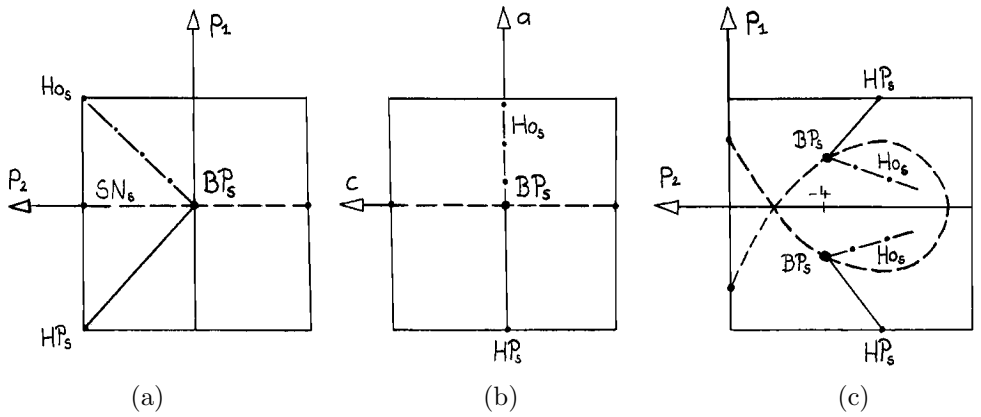


FIG. 2. *The singularly perturbed points at* $\epsilon = 0$: (a) *Standard,* (b) *FitzHugh–Nagumo,* (c) *van der Pol–Duffing.*

In the following theorem we collect the different types of singularly perturbed points that appear near a singularly perturbed and generically unfolded Bogdanov point.

THEOREM 2.3. *At* $\epsilon = 0$ *there exist in* (p_1, p_2)*-space of system* (2.5) *a two-sided curve* $SN_s(z)$, $-1 \ll z \ll 1$, $z \neq 0$ *of singularly perturbed saddle-nodes and two one-sided curves* $HP_s(z)$, $Ho_s(z)$, $0 < z \ll 1$ *of singularly perturbed Hopf points and singularly perturbed homoclinic base points, respectively. The curves can be smoothly continued to* $z = 0$. *At* $z = 0$ *the curves* $SN_s(z)$, $HP_s(z)$ *and* $SN_s(z)$, $Ho_s(z)$ *meet transversally in the singularly perturbed Bogdanov point at* $(p_1, p_2) = 0$. *The curves* $HP_s(z)$, $Ho_s(z)$ *exist on different sides of the curve* $SN_s(z)$ *and they meet transversally iff the Bogdanov point* $u = 0$ *of the system* $\dot{u} = F(u, 0)$ *is nondegenerate, i.e., if a certain normal form coefficient is different from zero.*

The theorem is proved in sections 4 and 5. We derive explicit formulas for the derivatives $SN_s{}'(0)$, $HP_s{}'(0)$, and $Ho_s{}'(0)$ in terms of derivatives of the right-hand side F and G.

In diagram (a) of Figure 2 we show the general location of the singularly perturbed points in the (p_1, p_2)-plane as they occur for the system (2.5) in generic cases.

The singularly perturbed Hopf points HP_s and the singularly perturbed homoclinic base points Ho_s terminate at the singularly perturbed Bogdanov point BP_s, whereas the singularly perturbed saddle-nodes SN_s pass through this point. The points HP_s and Ho_s occur on different sides of the curve of SN_s points. For graphical reasons the points SN_s are drawn along the p_2-axis.

In the case of the FitzHugh–Nagumo equation (1.3) we obtain diagram (b) in Figure 2; i.e., the singularly perturbed homoclinic orbits and the singularly perturbed Hopf points are positioned on the a-axis. In particular they do not meet transversally at the origin. This is a degenerate situation which results from the fact that one of the normal form coefficients of the central u-system $\dot{u}_1 = u_2$, $\dot{u}_2 = -u_1^2 + u_1^3$ vanishes.

Diagram (c) in Figure 2 shows the singularly perturbed points of the van der Pol–Duffing oscillator (1.2) which represents an example of two times the standard case. Further examples of the standard case are given by the model equation (1.5) and the cell communication problems studied in [25].

The singularly perturbed saddle-node and Hopf points in the three systems (1.5),

(1.3), and (1.2) can be given explicitly. In general it is only possible to calculate the derivatives $SN_s'(0)$, $HP_s'(0)$, and $Ho_s'(0)$ of the curves at the singularly perturbed Bogdanov point. The following are the parameter values of the singularly perturbed points. In the case of the model equation (1.5) we obtain

$$\begin{aligned} SN_s: &\quad (\lambda\ ,k) = (0\ ,k)\ , &&\quad k \neq 0,\\ HP_s: &\quad (\lambda,k)\ = (\lambda,-\lambda), &&\quad \lambda < 0,\\ Ho_s: &\quad (\lambda,k)\ = (\lambda,k(\lambda)), &&\quad \lambda > 0 \quad \text{with} \quad \tfrac{d}{d\lambda}k(\lambda) \to \tfrac{5}{7} \quad \text{as} \quad \lambda \to 0 \end{aligned}$$

(cf. diagram (a) in Figure 2 with $p_1 = \lambda$, $p_2 = k$). In the case of the FitzHugh–Nagumo equation (1.3) the singularly perturbed points occur at

$$\begin{aligned} SN_s: &\quad (a,c) = (0,c), &&\quad c \neq 0,\\ HP_s: &\quad (a,c) = (a,0), &&\quad a < 0,\\ Ho_s: &\quad (a,c) = (a,0), &&\quad a > 0 \end{aligned}$$

(cf. diagram (b) in Figure 2) and, finally, for the van der Pol–Duffing oscillator (1.2) we have

$$\begin{aligned} SN_s: &\quad (p_1,p_2) = (5z^3 - 3z, 6z^2 - 6), &&\quad z \in (-1,1) \setminus \{\pm\tfrac{1}{\sqrt{3}}\},\\ HP_s: &\quad (p_1,p_2) = (\pm(p_2 + \tfrac{8}{3})/\sqrt{3}, p_2), &&\quad p_2 < -4,\\ Ho_s: &\quad (p_1,p_2) = (p_1^{\mp}(p_2), p_2) &&\quad \text{with} \quad \tfrac{d}{dp_2}p_1^{\mp}(p_2) \to \mp\tfrac{1}{5\sqrt{3}} \quad \text{as} \quad p_2 \to -4 \end{aligned}$$

(cf. diagram (c) in Figure 2). The curve SN_s is globally parametrized by the external parameter z such that the singularly perturbed Bogdanov points BP_s occur at $z = \pm 1/\sqrt{3}$.

Next we perturb these basic structures from $\epsilon = 0$ to $\epsilon \neq 0$; i.e., we perform the actual singular perturbation. Concerning this process note that the (u,v)-linearization of the system (2.5) at the singularly perturbed Bogdanov point $(u,v,p,\epsilon) = 0$ reads

$$\begin{pmatrix} F_u^0 & F_v^0 \\ 0\cdot G_u^0 & 0\cdot G_v^0 \end{pmatrix} \in \mathbb{R}^{3,3} \quad \text{with} \quad \operatorname{tr}(F_u)^0 = 0 \quad \text{and} \quad \det(F_u)^0 = 0 \tag{2.9}$$

implying an eigenvalue zero of multiplicity three. The multiplicity three results from multiplicity two of the Bogdanov point and the degeneracy caused by the continuum of stationary points that exists in the (u,v)-phase space of system (2.5) at $\epsilon = 0$ (cf. Figure 1(b)). Hence without considering the singular perturbation character of the system (2.5) we are confronted with a highly degenerate situation. On the other hand, the assumption (2.4) implies

$$\det\left(\begin{pmatrix} F_u & F_v \\ \epsilon\cdot G_u & \epsilon\cdot G_v \end{pmatrix}\right)(u,v,p_1,p_2,\epsilon) = \epsilon \cdot \det\left(\begin{pmatrix} F_u & F_v \\ G_u & G_v \end{pmatrix}\right)(u,v,p_1,p_2,\epsilon) \neq 0 \tag{2.10}$$

for $\epsilon \neq 0$; i.e., the threefold eigenvalue zero splits completely into eigenvalues different from zero during variation from $\epsilon = 0$ to $\epsilon \neq 0$. In particular there exist no saddle-node or Bogdanov points for $\epsilon \neq 0$ which simplifies the unfolding problem considerably.

Our aim is to extend the (p_1, p_2)-unfolding diagrams of Figure 2 in the direction of ϵ. We use results [8], [11], [1], [10], [26], and [27] that are valid on 2-dimensional sections in the 3-dimensional (p_1, p_2, ϵ)-unfolding space. These sections are indicated in diagram (c) of Figure 1. The theorems describe (to some extent) the unfolding structure near the singularly perturbed saddle-node and Hopf points. As usual these 2-dimensional diagrams shrink to zero when approaching the more degenerate situation of a singularly perturbed Bogdanov point at $p = 0$.

On the contrary, the homoclinic orbits are investigated by blowing up the singularly perturbed Bogdanov point BP_s appropriately; i.e., here we do not restrict ourselves to 2-dimensional sections, but we start from the center of the 3-dimensional (p_1, p_2, ϵ)-unfolding space. Concerning this process we apply a theorem from [4]. It should be noted though that we could also use blow-up transformations for the investigation of the singularly perturbed saddle-node and Hopf points. However, these scaling transformations hardly showed further information, so we restricted ourselves to the theorems from [26], [27] that were applicable in a direct way.

A lot of gaps remain for a complete understanding of the (p_1, p_2, ϵ)-unfolding diagram. Some of these gaps are examined by a combination of numerical results and heuristical considerations mainly obtained from the model equation (1.5).

In this case we arrive at the unfolding structures depicted in diagram (a) of Figure 3 (use $p_1 = \lambda$, $p_2 = k$). The numerical calculations were performed with [9] and programs from [22].

Roughly speaking, the diagram is characterized by three surfaces: A surface of Hopf points HP, a surface of Naimark–Sacker points NP (hatched), and a surface of homoclinic orbits (or homoclinic base points) Ho. The surface of Hopf points connects the singularly perturbed saddle-nodes SN_s and the singularly perturbed Hopf points HP_s that exist at $\epsilon = 0$. In addition the points HP_s give rise to the surface NP of Naimark–Sacker points; i.e., we obtain a surface of periodic orbits with a pair of complex-conjugate Floquet multipliers on the unit circle which implies generically the existence of a 3-dimensional domain of invariant tori near the surface of Naimark–Sacker points (cf. [15]).

The surface NP terminates at the curve NP_{-1} of periodic orbits with double Floquet multiplier -1. On the other hand we obtain the threefold Floquet multiplier 1 along the curve of HP_s-points at $\epsilon = 0$. Hence during variation along the dotted line [a] the two Floquet multipliers different from 1 pass along the unit circle from 1 to -1 as indicated in diagram (b). This leads to curves of NP-points in diagram (a) with fixed Floquet multipliers on the unit circle. These curves approach the central point $(p_1, p_2, \epsilon) = 0$. As an example, the curve of NP-points with fixed Floquet multipliers $\pm i$ is indicated by NP_i.

Passing to the study of maps via a Poincaré section the points NP_{-1} and NP_i are known as $1:2$ and $1:4$ resonances, respectively (see [20] for their unfolding). We conclude that the central (u, v)-system

$$\begin{aligned} \dot{u} &= F(u, v, 0, 0, 0), \\ \dot{v} &= 0 \end{aligned} \tag{2.11}$$

at $(p_1, p_2, \epsilon) = 0$ gives rise to completely different kinds of periodic orbits. The

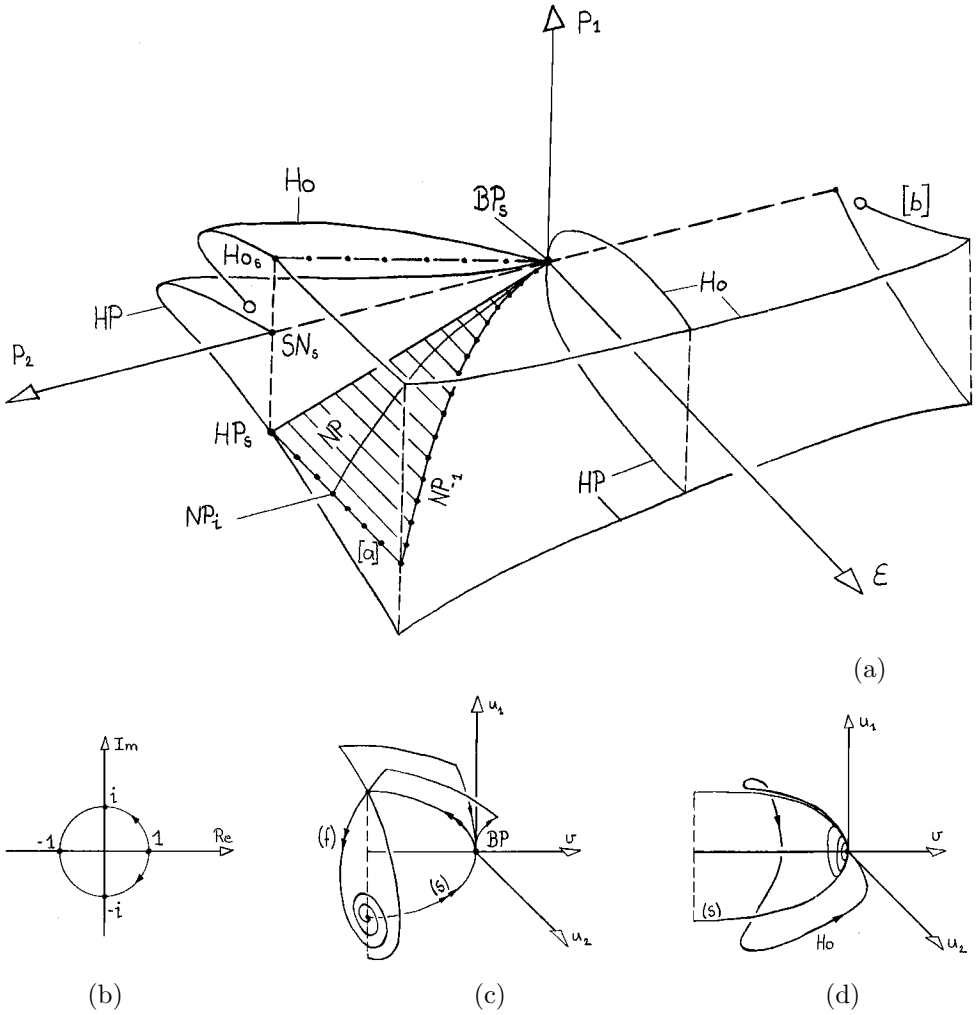


FIG. 3. (a) (p_1, p_2, ϵ)-*unfolding diagram;* (b) *Floquet multipliers along* [a]*;* (c) *a singular periodic orbit in the* (u, v)*-phase space of the central system* (2.11)*;* (d) *a typical homoclinic orbit of Shilnikov type near the border line of open circles.*

complete picture is unknown. Our suggestion is that the periodic orbits along NP_i or NP_{-1} start in a canard like manner from a singular periodic orbit (cf. [10]) of the system (2.11) composed of fast (f) and slow motions (s) as indicated by double arrows in the (u, v)-phase diagram (c). Note that the central (u, v)-system (2.11) inherits the continuum of stationary points (s) which consists of saddles and oscillating sinks (with respect to the fast u-dynamics) separated by a Bogdanov point BP at $v = 0$. Note also that diagram (c) is identical to the bifurcation diagram of the fast u-system shown in the lower half of diagram (b) in Figure 1.

Next we turn to the surface of homoclinic orbits Ho in diagram (a) that bifurcates from the singularly perturbed homoclinic base points Ho_s at $\epsilon = 0$. In general this bifurcation is not smooth in ϵ, i.e., the surface has an edge along the curve of Ho_s-points at $\epsilon = 0$ (cf. section 5).

Further, as indicated by the open circles the surface Ho is bordered by a curve

through the central point BP_s along which the homoclinic orbits vanish presumably by a canard like blow-up. More precisely, along the curve [b] we expect that the homoclinic orbits turn first into homoclinic orbits of Shilnikov type [15] and finally blow-up along the continuum of stationary points (s) of the central system (2.11) as indicated in diagram (d). Some additional features of this explosion process are described in section 5.6.

Now we return to the general system (2.5) under the assumptions (2.2)–(2.4) and (2.7). We expect that the structures in Figure 3 of the model equation (1.5) represent a typical unfolding diagram of singularly perturbed Bogdanov points in a generic situation. Some aspects of this claim are verified analytically in the next sections. Moreover, numerical calculations for the van der Pol–Duffing oscillator (1.2) (cf. [19]) and the cell communication problems in [25] confirm this conjecture.

We notice, however, that the Hamiltonian case (1.3) of the FitzHugh–Nagumo equation exhibits some special features (cf. diagram (b) in Figure 2). Therefore, we close this section with some remarks on deformations of the unfolding picture that might occur in specific examples.

First, the surfaces in Figure 3 diagram (a) may be bended diffeomorphically as usual; i.e., we are only interested in structures that are not destroyed by diffeomorphic parameter transformations. For example, the folds of Hopf and homoclinic base points shown in diagram (a) may vanish under an appropriate change of the parameters. On the other hand the transversal intersections of manifolds of singularly perturbed points represent diffeomorphic invariants.

Second, a singular perturbation problem of the form (2.5) contains at $(p_1, p_2, \epsilon) = 0$ a continuum of stationary points in the (u, v)-phase space defined by $F(u, v, 0) = 0$ (cf. (s) in diagram (c) of Figure 3). Even in the weak sense of algebraic singularity theory [13] this leads to a completely degenerate situation, i.e., a codim $= \infty$ situation occurs, so that we cannot hope to grasp all possible phenomena near a singularly perturbed Bogdanov point with the help of a certain normal form. In this sense singular perturbation problems do not fit into the framework of universal unfoldings which are based on the fact that higher order derivatives do not change the qualitative behavior of the system.

Nevertheless, every singularly perturbed Bogdanov point with a generic unfolding will show the properties that are completely determined by the derivatives involved in (2.2)–(2.4) and (2.7). But the phenomena which depend on higher derivatives will change for different singularly perturbed Bogdanov points. In spite of this principal lack of completeness which occurs in singular perturbation problems we hope that our results grasp the essential features of singularly perturbed Bogdanov points.

3. A pretransformation. First we perform a pretransformation of the system (2.5) under the slightly weaker assumption (cf. (2.7))

$$\text{rank} \left(\begin{pmatrix} F_u & F_v & F_{p_1} & F_{p_2} \\ G_u & G_v & G_{p_1} & G_{p_2} \\ \det_u(F_u) & \det_v(F_u) & \det_{p_1}(F_u) & \det_{p_2}(F_u) \end{pmatrix}^0 \right) = 4. \tag{3.1}$$

The transformed system will represent a certain normal form of a singularly perturbed Bogdanov point within a generic unfolding.

According to the assumption (2.2) the point $u = 0$ is a Bogdanov point of the system $\dot{u} = F(u, 0)$. Hence there exist $\varphi_2, \psi_1, \psi_2 \in \mathbb{R}^2 \setminus \{0\}$ with

$$F_u^0 \varphi_2 = \varphi_1, \qquad \psi_2^T F_u^0 = 0, \qquad \psi_1^T F_u^0 = \psi_2^T, \qquad \psi_i^T \varphi_j = \delta_{ij}. \tag{3.2}$$

Further, the system $\dot{u} = F(u, v, 0)$ has a quadratic limit point with respect to v (cf. (2.3)). Thus, there exist a surface of quadratic limit points $(u, v)^L(p_1, p_2, \epsilon)$ of $\dot{u} = F(u, v, p_1, p_2, \epsilon)$ and smooth tangent vectors $\phi_1(p_1, p_2, \epsilon)$ such that $(u, v)^L(0) = 0$, $\phi_1(0) = \varphi_1$. This follows from the implicit function theorem applied to the equation $[F, \det(F_u)][u, v, p_1, p_2, \epsilon] = 0$. We obtain the identity $[F, \det(F_u)^0][(u, v)^L(p_1, p_2, \epsilon), p_1, p_2, \epsilon] = 0$ and differentiation with respect to $p = (p_1, p_2)$ yields

$$\begin{pmatrix} u_p^L \\ v_p^L \end{pmatrix}(0) = -\begin{pmatrix} F_u^0 & F_v^0 \\ \det_u(F_u)^0 & \det_v(F_u)^0 \end{pmatrix}^{-1} \cdot \begin{pmatrix} F_p^0 \\ \det_p(F_u)^0 \end{pmatrix}, \tag{3.3}$$

where the regularity of the matrix follows from (2.2), (2.3). Next we substitute the limit points into the equation $G = 0$; i.e., we analyze $G[(u, v)^L(p, \epsilon), p, \epsilon] = 0$. At $(p, \epsilon) = 0$ we have $G[(u, v)^L(0), 0] = G^0 = 0$ and (3.1), (3.3) imply

$$\frac{d}{dp} G[(u, v)^L(0), 0] = (G_u^0, G_v^0) \cdot \begin{pmatrix} u_p^L \\ v_p^L \end{pmatrix}(0) + G_p^0 \neq (0, 0). \tag{3.4}$$

Without loss of generality we may assume $\frac{d}{dp_1} G[(u, v)^L(0), 0] \neq 0$. Hence there exists a unique smooth function $p_1(p_2)$ satisfying $p_1(0) = 0$ and

$$G[(u, v)^L(p_1(p_2), p_2, 0), p_1(p_2), p_2, 0] = 0; \tag{3.5}$$

i.e., within the 2-dimensional surface of limit points $(u, v)^L(p, 0)$ of the u-system $\dot{u} = F(u, v, p, 0)$ we obtain a curve of limit points which satisfies $G = 0$. Now we define the matrices $\phi(p, \epsilon) := [\phi_1(p, \epsilon), \varphi_2] \in \mathbb{R}^{2,2}$ and consider the transformation

$$\begin{aligned} p_1 &= p_1(k) + \lambda, \\ p_2 &= k, \\ u &= u^L[p_1(k) + \lambda, k, \epsilon] + \phi[p_1(k) + \lambda, k, \epsilon] \cdot x =: u(x, \lambda, k, \epsilon), \\ v &= v^L[p_1(k) + \lambda, k, \epsilon] + y =: v(y, \lambda, k, \epsilon); \end{aligned} \tag{3.6}$$

i.e., the limit points are moved to the origin of phase space, the limit points which satisfy $G = 0$ at $\epsilon = 0$ are positioned on the k-axis, and a linear normalization of the u-components is performed. We arrive at the system

$$\begin{aligned} \dot{x} &= \phi^{-1}[p_1(k) + \lambda, k, \epsilon] \cdot F[u(x, \lambda, k, \epsilon), v(y, \lambda, k, \epsilon), p_1(k) + \lambda, k, \epsilon] \\ &=: f(x, y, \lambda, k, \epsilon), \\ \dot{y} &= \epsilon \cdot G[u(x, \lambda, k, \epsilon), v(y, \lambda, k, \epsilon), p_1(k) + \lambda, k, \epsilon] \\ &=: \epsilon \cdot g(x, y, \lambda, k, \epsilon) \end{aligned} \tag{3.7}$$

which satisfies

$$f(0, 0, \lambda, k, \epsilon) \equiv 0, \qquad f_{x_1}(0, 0, \lambda, k, \epsilon) \equiv 0, \qquad g(0, 0, 0, k, 0) \equiv 0,$$

$$\begin{pmatrix} f_x & f_y \\ g_x & g_y \end{pmatrix}^0 = \begin{pmatrix} 0 & 1 & \psi_1^T F_v^0 \\ 0 & 0 & \psi_2^T F_v^0 \\ G_u^0 \varphi_1 & G_u^0 \varphi_2 & G_v^0 \end{pmatrix} \quad \text{nonsingular,}$$

$$g_\lambda^0 = (G_u \quad G_v \quad G_{p_1})^0 \cdot \begin{pmatrix} u_{p_1}^L \\ v_{p_1}^L \\ 1 \end{pmatrix}(0) \neq 0,$$

(3.8)

$$p_1'(0) = -\frac{1}{g_\lambda^0}(G_u \quad G_v \quad G_{p_2})^0 \cdot \begin{pmatrix} u_{p_2}^L \\ v_{p_2}^L \\ 1 \end{pmatrix}(0),$$

$$(f_{x_1x_1} \quad f_{x_1x_2})^0 = \begin{pmatrix} \psi_1^T \\ \psi_2^T \end{pmatrix} \cdot (F_{uu}^0\varphi_1^2 \quad F_{uu}^0\varphi_1\varphi_2) \qquad \text{with} \qquad f_{2x_1x_1}^0 \neq 0,$$

$$(f_{2x_2\lambda} \quad f_{2x_2k})^0 = (\mathrm{tr}_u(F_u) \quad \mathrm{tr}_v(F_u) \quad \mathrm{tr}_p(F_u))^0 \cdot \begin{pmatrix} u_p^L \\ v_p^L \\ I_2 \end{pmatrix}(0) \cdot \begin{pmatrix} 1 & p_1'(0) \\ 0 & 1 \end{pmatrix}.$$

The last identity follows after a lengthy but straightforward calculation using (3.2) and

$$F_u[(u,v)^L(p,0),p,0]\cdot\phi_1(p,0) = 0,$$

(3.9)

$$\mathrm{tr}_a(F_u)^0 = \psi_1^T F_{ua}^0\varphi_1 + \psi_2^T F_{ua}^0\varphi_2 \qquad \text{for} \qquad a = u_1, u_2, v, p_1, p_2.$$

Moreover we obtain from (3.3) and (3.8) after some calculations

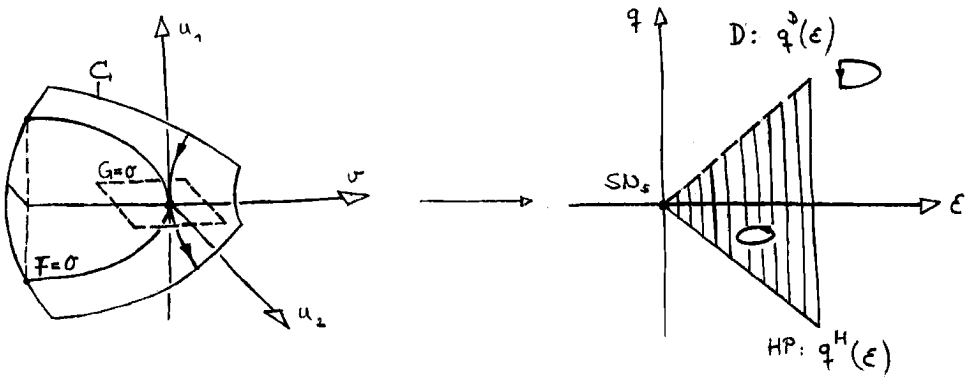
(3.10)

$$\begin{aligned} f_{2x_2k}^0 &= -\frac{1}{g_\lambda^0}\cdot\det\left(\begin{pmatrix}\mathrm{tr}_u(F_u) & \mathrm{tr}_v(F_u)\\ G_u & G_v\end{pmatrix}^0\cdot\begin{pmatrix}u_p^L\\ v_p^L\end{pmatrix}(0) + \begin{pmatrix}\mathrm{tr}_p(F_u)\\ G_p\end{pmatrix}^0\right)\\ &= \frac{1}{g_\lambda^0}\cdot\det\left(\begin{pmatrix}\mathrm{tr}_u(F_u) & \mathrm{tr}_v(F_u)\\ G_u & G_v\end{pmatrix}\begin{pmatrix}F_u & F_v\\ \det_u(F_u) & \det_v(F_u)\end{pmatrix}^{-1}\begin{pmatrix}F_p\\ \det_p(F_u)\end{pmatrix}\right.\\ &\quad \left. - \begin{pmatrix}\mathrm{tr}_p(F_u)\\ G_p\end{pmatrix}\right)^0. \end{aligned}$$

Note that the defining condition (2.7) of a generic unfolding implies $f_{2x_2k}^0 \neq 0$, whereas the weaker assumption (3.1) of this section does not exclude the case $f_{2x_2k}^0 = 0$. The homoclinic orbits of the systems (2.5) and (3.7) depend heavily on this coefficient. In the analysis of singularly perturbed saddle-node and Hopf points (cf. section 4) we restrict to $f_{2x_2k}^0 \neq 0$, whereas the singularly perturbed homoclinic orbits are investigated for $f_{2x_2k}^0 = 0$ as well (cf. section 5). In the case of $f_{2x_2k}^0 = 0$ we require further coefficients to be different from zero.

4. Unfolding singularly perturbed saddle-node and Hopf points. First we state the necessary definitions. This is done completely analogous to Definitions 2.1 and 2.2 of a singularly perturbed Bogdanov point and its generic unfolding. For definiteness the singularly perturbed points are assumed to lie in the origin.

DEFINITION 4.1. *A point* $(u,v,\epsilon) = 0$ *which satisfies* $tr(F_u)^0 \neq 0$ *and the remaining conditions in* (2.2)–(2.4) *is called a singularly perturbed saddle-node of the system* (2.1).



(a) (b)

FIG. 4. (a) *The basic configuration for the fast subsystem* $\dot{u} = F(u, v, 0, 0)$ *and the surface defined by* $G(u, v, 0, 0) = 0$, *C=center manifold;* (b) *the corresponding* (q, ϵ)*-unfolding diagram of the singularly perturbed saddle-node* SN_s.

DEFINITION 4.2. *A point* $(u, v, \epsilon) = 0$ *which satisfies* $det(F_u)^0 > 0$ *and the remaining conditions in* (2.2) *and* (2.4) *is called a singularly perturbed Hopf point of the system* (2.1).

Apart from the basic condition $(F, G)^0 = 0$ singularly perturbed saddle-node and Hopf points satisfy the degeneracy condition $\det(F_u)^0 = 0$ and $\mathrm{tr}(F_u)^0 = 0$, respectively. Hence we are led to investigate these points within a parameter unfolding of one parameter $q \in \mathbb{R}$

$$\begin{aligned} \dot{u} &= F(u, v, q, \epsilon) \in \mathbb{R}^2, \\ \dot{v} &= \epsilon \cdot G(u, v, q, \epsilon) \in \mathbb{R}. \end{aligned} \tag{4.1}$$

The corresponding nondegeneracy condition of the parameter unfolding (4.1) is again determined with the help of the defining equations

$$H_{SN_s}(u, v, q) := \begin{pmatrix} F(u, v, q, 0) \\ G(u, v, q, 0) \\ \det(F_u)(u, v, q, 0) \end{pmatrix} = 0, \qquad H_{SN_s} : \mathbb{R}^4 \to \mathbb{R}^4 \tag{4.2}$$

and

$$H_{HP_s}(u, v, q) := \begin{pmatrix} F(u, v, q, 0) \\ G(u, v, q, 0) \\ \mathrm{tr}(F_u)(u, v, q, 0) \end{pmatrix} = 0, \qquad H_{HP_s} : \mathbb{R}^4 \to \mathbb{R}^4, \tag{4.3}$$

respectively.

DEFINITION 4.3. *The q-unfolding* (4.1) *of system* (2.1) *is a generic unfolding of a singularly perturbed saddle-node (singularly perturbed Hopf point) if* $H'_{SN_s}(0) \in \mathbb{R}^{4,4}$ *is nonsingular (*$H'_{HP_s}(0) \in \mathbb{R}^{4,4}$ *is nonsingular).*

Singularly perturbed saddle-nodes have been examined in many papers; see, for example, [8], [11], [1], [10], [26]. We collect some results that are valid for a singularly perturbed saddle-node within a generic unfolding in the following theorem.

THEOREM 4.4. *There exists a* (q,ϵ)*-unfolding diagram of system* (4.1) *as depicted in diagram* (b) *of Figure* 4.

There exists a one-sided curve of Hopf points $q^H(\epsilon)$ *from which small periodic orbits bifurcate in the direction of* q*. The periodic orbits exist in the hatched region and they blow-up to some global object within an exponentially small strip of canard orbits indicated by the line* $q^D(\epsilon)$*.*

In diagram (a) we show the basic configuration for the fast subsystem $\dot{u} = F(u,v,0,0)$, $\dot{v} = 0$ with the saddle-node at $v = 0$. Here we indicate also an invariant center manifold C that comprises the continuum of stationary points (defined by $F(u,v,0,0) = 0$) and the saddle-node dynamics at $v = 0$. In [26] the precise location of $(q^H, q^D)(\epsilon)$ and the stability of the periodic orbits are expressed in terms of the reduced system on the center manifold. In the next section we repeat these calculations (to some extent) for the singularly perturbed saddle-nodes of the pretransformed system (3.7). Using the center manifold C Theorem 4.4 can be extended to $(n+1)$-dimensional systems of the form $\dot{x} = F(u,v,q,\epsilon) \in \mathbb{R}^n$, $\dot{v} = \epsilon G(u,v,q,\epsilon) \in \mathbb{R}$.

Next we turn to singularly perturbed Hopf points within a generic unfolding. In [27] points of this type are investigated under the additional nondegeneracy condition

$$\operatorname{tr}_u(F_u)^0 \cdot (F_u^0)^{-1} \cdot F_v^0 - \operatorname{tr}_v(F_u)^0 \neq 0 \ . \tag{4.4}$$

Comparing Definitions 4.1 and 4.2 of singularly perturbed saddle-node and Hopf points we see that condition (2.3) is skipped in definition 4.2 of a singularly perturbed Hopf point. Instead we assume now the nondegeneracy condition (4.4).

Both of the conditions (2.3) and (4.4) ensure that an eigenvalue of the fast u-system $\dot{u} = F(u,v,0,0)$ passes the imaginary axis with nonvanishing velocity along the branch of stationary points given by $F(u,v,0,0) = 0$. In the case of a singularly perturbed saddle-node the imaginary axis is passed at zero yielding a quadratic limit point with respect to v (cf. Figure 4 (a)). In case of a singularly perturbed Hopf point with (4.4) the imaginary axis is passed at $\pm i\sqrt{\det(F_u)^0}$ and we obtain the well-known Hopf bifurcation of periodic orbits for the fast u-system as indicated in diagram (a) of Figure 5 below. Before turning to the corresponding (q,ϵ)-unfolding diagram shown in Figure 5(b) we add some remarks concerning the necessity of the conditions (2.3) and (4.4).

In [26] the velocity condition (2.3) is ignored and we require further coefficients to be different from zero. We obtain singularly perturbed points of higher degeneracy, e.g., *singularly perturbed cusp points*. We believe that we could also relax the velocity condition (4.4) of a singularly perturbed Hopf point. This is motivated by numerical calculations and the fact that also the singularly perturbed homoclinic orbits treated in section 5 do not require a velocity condition with respect to v (intersection of stable and unstable manifold with nonvanishing velocity).

The condition (4.4) prevents the investigation of the singularly perturbed Hopf points of the FitzHugh–Nagumo equation (1.3). They occur at $c = 0$, $a < 0$ (cf. Figure 2(b)) within the family of Hamiltonian systems $\dot{u}_1 = u_2$, $\dot{u}_2 = u_1(u_1 - a)(u_1 - 1) + v$. We obtain for fixed $a < 0$ a v-dependent system that possesses a branch of stationary points which is completely filled up with Hopf points. Hence the eigenvalues remain on the imaginary axis along $F = 0$ and the results of this paper about singularly perturbed Hopf points (see Theorem 4.5 below) are not applicable in a direct way. However, note that the parameter values $a < 0$ are of no biological relevance. The results of this paper about singularly perturbed saddle-nodes and singularly perturbed homoclinic orbits are applicable to the FitzHugh–Nagumo equation as well. These phenomena occur for the biological relevant parameter values $a \geq 0$.

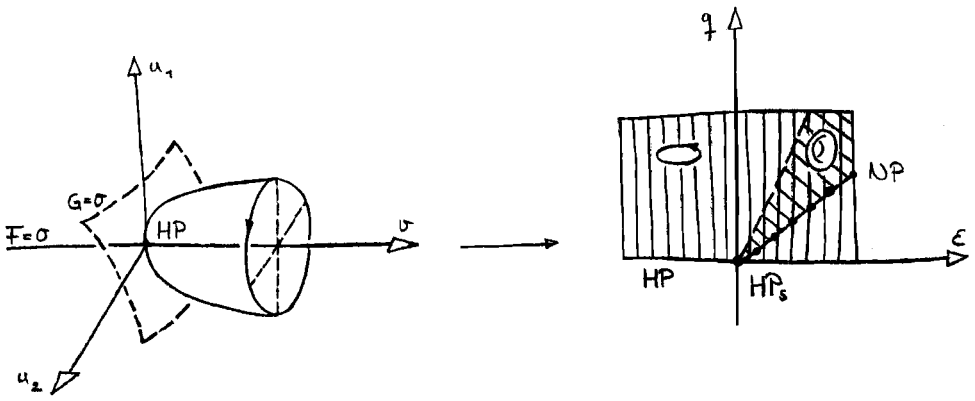


FIG. 5. (a) *The basic configuration for the fast subsystem* $\dot{u} = F(u, v, 0, 0)$ *and the surface defined by* $G(u, v, 0, 0) = 0$*;* (b) *a typical* (q, ϵ)*-unfolding diagram of a singularly perturbed Hopf point* HP_s*.*

We collect some results from [27] that are valid for a singularly perturbed Hopf point within a generic unfolding in the following theorem.

THEOREM 4.5. *Under the nondegeneracy condition* (4.4) *there exists a* (q, ϵ)*-unfolding diagram of system* (4.1) *as depicted in diagram* (b) *of Figure* 5*.*

There exists a curve of Hopf points HP through the origin (transversal to the q-axis) from which periodic orbits bifurcate in the direction of q. The periodic orbits exist in the weakly hatched region which represents a 2*-dimensional surface of periodic orbits. Depending on the sign of a characteristic constant* κ*, the surface contains a curve NP of periodic orbits with a conjugate-complex pair of Floquet multipliers on the unit circle. Typically these orbits give rise to invariant tori that are indicated in the cross hatched region.*

In the case of the van der Pol–Duffing oscillator the singularly perturbed Hopf points occur at $p_1 = \pm(p_2 + \frac{8}{3})/\sqrt{3}$, $p_2 < -4$. A corresponding unfolding diagram can be found in [19, Figure 12(a)] .

The rather technical derivation and an explicit formula of the characteristic constant κ can be found in [25], [27]. Figure 5(b) represents the case $\kappa < 0$. The precise properties of the cross hatched region with invariant tori are not known and its border is indicated by a broken line and drawn in an arbitrary fashion.

Note that at $\epsilon = 0$ not any invariant torus can survive in the resulting 2-dimensional system $\dot{u} = F(u, v, q, 0)$, $\dot{v} = 0$. This is a typical feature of singular perturbation problems, where the lower dimensional $\epsilon = 0$ system cannot show the higher dimensional phenomena that may exist in the full system. Note also that the invariant tori appear as a consequence of coupling effects between the fast x-variables and the slow y-variable. Analogous to singularly perturbed saddle-nodes we can extend Theorem 4.5 to $(n + 1)$-dimensional systems of the form $\dot{u} = F(u, v, q, \epsilon) \in \mathbb{R}^n$, $\dot{v} = \epsilon \cdot G(u, v, q, \epsilon) \in \mathbb{R}$ with the help of a reduction to a center manifold.

In concrete examples it is often rather cumbersome to calculate explicitly all the characteristic coefficients that determine quantitatively the unfolding structure depicted in diagrams (b) of Figures 4 and 5 (stability, bifurcation direction, ...). In particular at a singularly perturbed Hopf point higher order variational equations

must be solved which suggests using computeralgebraic programs. In the case of singularly perturbed Hopf points we shall mainly restrict to numerical results.

In [25], [26] we extend our unfolding scheme of singularly perturbed points to singularly perturbed periodic orbits using an averaging procedure. Additionally, we relax in [25] the transversality condition (2.4) for the nullclines of F and G. In a certain sense we obtain a hierarchical structure of singular perturbation problems in case the slow manifold loses its hyperbolicity which means that the well-known reduction to the slow manifold [12] is not possible.

In the next section we prove first the existence and location of the curves of singularly perturbed saddle-nodes SN_s and Hopf points HP_s as depicted in Figure 2. Then we shall see that these points are generically unfolded and we can apply Theorems 4.4 and 4.5 on appropriate sections in the (p_1, p_2, ϵ)-unfolding space of a singularly perturbed Bogdanov point.

This is done for the pretransformed (x, y)-system (3.7). It is always obvious how to transfer the results from the (λ, k, ϵ)-space into the (p_1, p_2, ϵ)-parameter space of the original system (2.5). The p_1-axis and the p_2-axis in Figures 2 and 3 are often referred to as λ-axis and k-axis, respectively.

4.1. Singularly perturbed saddle-nodes. Recalling (3.8), (3.10) and Definitions 4.1 and 4.3 it is straightforward to see that the points $(x, y, \lambda, k, \epsilon) = (0, 0, 0, k, 0)$ represent for fixed $k \neq 0$ singularly perturbed saddle-nodes of system (3.7) within a generic $(q, \epsilon) = (\lambda, \epsilon)$-unfolding. Note that the linearization at these points is given by

$$\begin{pmatrix} f_x & f_y \\ 0 \cdot g_x & 0 \cdot g_y \end{pmatrix}(0, k, 0) = \begin{pmatrix} 0 & f_{1x_2} & f_{1y} \\ 0 & f_{2x_2} & f_{2y} \\ 0 & 0 & 0 \end{pmatrix}(0, k, 0) \tag{4.5}$$

with $f_{1x_2}(0, k, 0) = 1 + O(k)$, $f_{2x_2}(0, k, 0) = f^0_{2x_2k} \cdot k + O(k^2)$, $f_{2y}(0, k, 0) \neq 0$, and $f^0_{2x_2k} \neq 0$ due to the assumption of a generic unfolding of a singularly perturbed Bogdanov point.

We obtain for every $k \neq 0$ the algebraically double and geometrically simple eigenvalue zero with generalized eigenspace $\operatorname{span}\{(1, 0, 0)^T, (0, f_{2y}(0, k, 0), -f_{2x_2}(0, k, 0))^T\}$ and hence there exists for every $k \neq 0$ a local center manifold which can be parametrized by the coordinates (x_1, y) and the parameters (λ, k, ϵ) (cf. [24] and Figure 4 (a)). Strictly speaking there exists a smooth function $x_2(x_1, y, \lambda, k, \epsilon)$, $k \neq 0$ such that the surface

$$(x_1, x_2, y) = (x_1, x_2(x_1, y, \lambda, k, \epsilon), y), \qquad k \neq 0 \tag{4.6}$$

remains invariant under the flow of system (3.7) in the (x_1, x_2, y)-phase space. Graphically, we obtain for fixed $k \neq 0$ and $(\lambda, \epsilon) = 0$ an invariant manifold C as indicated in diagram (a) of Figure 4. The x_1-expansion at $(y, \lambda, \epsilon) = 0$ is given by

$$x_2(x_1, 0, 0, k, 0) = -\frac{1}{2}\frac{f_{2x_1x_1}}{f_{2x_2}}(0, k, 0) \cdot x_1^2 + O(x_1^3), \qquad k \neq 0. \tag{4.7}$$

The dynamics on the invariant surface reads

$$\begin{aligned} \dot{x}_1 &= f_1(x_1, x_2(x_1, y, \lambda, k, \epsilon), y, \lambda, k, \epsilon) =: \bar{f}(x_1, y, \lambda, k, \epsilon), \\ \dot{y} &= \epsilon \cdot g(x_1, x_2(x_1, y, \lambda, k, \epsilon), y, \lambda, k, \epsilon) =: \epsilon \cdot \bar{g}(x_1, y, \lambda, k, \epsilon) \end{aligned} \tag{4.8}$$

and we again have the ordinary singular perturbation form (1.1). At $(x_1, y, \lambda, k) = (0, 0, 0, k)$ we obtain quadratic limit points with respect to y of the x_1-system $\dot{x}_1 = \bar{f}(x_1, y, 0, k, 0)$ satisfying $\bar{g}(0, 0, 0, k, 0) = 0$. Along the section $y = 0$ the fast x_1-dynamics on the center manifold is given by

$$\begin{aligned} \dot{x}_1 &= \bar{f}(x_1, 0, 0, k, 0) \\ &= -\frac{1}{2f_{2x_2}(0,k,0)} \cdot \det\left(\begin{pmatrix} f_{1x_2} & f_{1x_1x_1} \\ f_{2x_2} & f_{2x_1x_1} \end{pmatrix}\right)(0,k,0) \cdot x_1^2 + O(x_1^3) \end{aligned} \tag{4.9}$$

as can be seen by direct calculation. Recalling (3.8) the leading coefficient in (4.9) is well defined and different from zero for $k \neq 0$ which implies a saddle-node dynamics on the center manifold [15]. Qualitatively, this saddle-node dynamics is depicted in Figure 4(a) along the section $v = 0$. As expected the leading coefficient in (4.9) tends to infinity if we approach the Bogdanov point at $k = 0$.

We can apply Theorem 4.4 to the singularly perturbed saddle-nodes $(x, y, \lambda, k, \epsilon) = (0, 0, 0, k, 0)$, $k \neq 0$ of system (3.7), and obtain (λ, ϵ)-unfolding diagrams as depicted in diagram (b) of Figure 4.

The precise location of $(q^H, q^D)(\epsilon, k) = (\lambda^H, \lambda^D)(\epsilon, k)$ and the stability of the periodic orbits depend on certain derivatives of $\bar{f}$ and $\bar{g}$ at the singularly perturbed saddle-nodes. With the setting $(\bar{f}, \bar{g})^0 := (\bar{f}, \bar{g})(0, k, 0)$, $k \neq 0$ and with $-\bar{g}^0_{x_1}\bar{f}^0_y > 0$, it is shown in [25], [26] that the hatched region of periodic orbits exists for $\epsilon > 0$ and we obtain

$$\lambda^H_\epsilon(0,k) = -\frac{A+G}{B} \qquad \text{and} \qquad \lambda^D_\epsilon(0,k) = -\frac{(C+F)E+3D}{BE^2}, \tag{4.10}$$

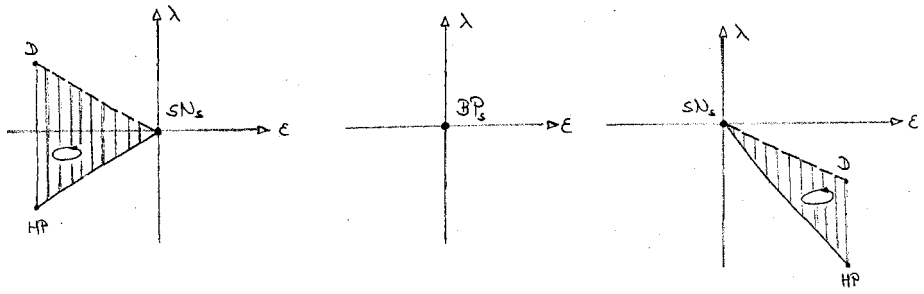
where

$$\begin{aligned} d &:= \sqrt{-\bar{g}^0_{x_1}\bar{f}^0_y} > 0, \\ A &:= \tfrac{1}{d^3}\cdot(\bar{f}_{x_1x_1}(\bar{f}_y\bar{g}_\epsilon - \bar{g}_y\bar{f}_\epsilon) - \bar{g}_{x_1}(\bar{f}_{\epsilon x_1}\bar{f}_y - \bar{f}_{yx_1}\bar{f}_\epsilon))^0, \\ B &:= \tfrac{1}{d^3}\cdot(\bar{f}_{x_1x_1}(\bar{f}_y\bar{g}_\lambda - \bar{g}_y\bar{f}_\lambda) - \bar{g}_{x_1}(\bar{f}_{\lambda x_1}\bar{f}_y - \bar{f}_{yx_1}\bar{f}_\lambda))^0 \neq 0, \\ C &:= \bar{f}^0_{x_1y}, \\ D &:= \tfrac{1}{6d}\cdot(\bar{f}_{x_1x_1x_1}\bar{f}^2_y - 3\bar{f}_{x_1y}\bar{f}_{x_1x_1}\bar{f}_y)^0, \\ E &:= \tfrac{1}{d}\cdot(\bar{f}_{x_1x_1}\bar{f}_y)^0 \neq 0, \\ F &:= \tfrac{1}{2d^2}\cdot(\bar{g}_{x_1x_1}\bar{f}^2_y - \bar{g}_y\bar{f}_{x_1x_1}\bar{f}_y)^0, \\ G &:= \tfrac{1}{d}\cdot\bar{g}^0_y\ . \end{aligned} \tag{4.11}$$

The inequalities $-\bar{g}^0_{x_1}$, $\bar{f}^0_y \neq 0$ and B, $E \neq 0$ follow by direct calculation from (3.8). In the case of $-\bar{g}^0_{x_1}\bar{f}^0_y < 0$ the periodic orbits exist for $\epsilon < 0$ and the formulas (4.10), (4.11) can be applied after performing the trivial transformation $\epsilon \to -\epsilon$.

Now the model equation (1.5) yields for $k > 0$ and $k < 0$ the left and right diagram in Figure 6, respectively.

For $k > (<)\ 0$ the periodic orbits exist locally for $\epsilon < (>)\ 0$. The derivative of the curve of Hopf points at $\epsilon = 0$ reads $\lambda^H_\epsilon(0,k) = \frac{1}{2k}$ whereas the canard orbits satisfy $\lambda^D_\epsilon(0,k) \equiv -\frac{1}{4}$. Note that our center manifold approach breaks down as



$k > 0$ $k = 0$ $k < 0$

FIG. 6. *The (λ, ϵ)-unfolding diagrams near the singularly perturbed saddle-nodes of the model equation* (1.5).

k tends to zero and we obtain no information in the middle diagram at $k = 0$. The diagrams represent incomplete (λ, ϵ)-sections through the $(p_1, p_2, \epsilon) = (\lambda, k, \epsilon)$-unfolding diagram in Figure 3(a) near the singularly perturbed saddle-nodes SN_s. Similar results are valid for the general system (2.5) and hence for the FitzHugh–Nagumo equation (1.3) and the van der Pol–Duffing oscillator (1.2). The formulas become much more involved and we do not state them here. Note that the derivatives in (4.11) are calculated after the center manifold reduction is performed.

4.2. Singularly perturbed Hopf points. In this section we are looking for Hopf points of the fast x-system $\dot{x} = f(x, y, \lambda, k, 0)$ from (3.7) which satisfy $g(x, y, \lambda, k, 0) = 0$. The corresponding defining equation reads

$$\begin{pmatrix} f \\ g \\ \operatorname{tr}(f_x) \end{pmatrix} (x, y, \lambda, k, 0) =: H(x, y, k, \lambda) = 0, \qquad H : \mathbb{R}^5 \to \mathbb{R}^4. \tag{4.12}$$

From (3.8) we obtain $H(0) = 0$ and $H_{x,y,k}(0)$ regular. Hence there exists a unique function $(x, y, k)^H(\lambda)$ with $(x, y, k)^H(0) = 0$ and $H[(x, y, k)^H(\lambda), \lambda] = 0$. Moreover, with the setting $\bar{H}(\lambda) := \det(f_x)[(x, y)^H(\lambda), \lambda, k^H(\lambda), 0]$ a simple calculation yields $\bar{H}'(0) = f^0_{2x_1x_1} g^0_\lambda / g^0_{x_1} \neq 0$; i.e., in the case of $\bar{H}'(0) > (<)\ 0$ the curve $(x, y, k)^H(\lambda)$ represents for $\lambda > (<)\ 0$ a curve of Hopf points in the family of x-systems $\dot{x} = f(x, y, \lambda, k, 0)$. Moreover we have $g = 0$ along this curve and hence a curve of singularly perturbed Hopf points within a generic $(q, \epsilon) = (k, \epsilon)$-unfolding occurs. The derivative of the k-component at $\lambda = 0$ reads

$$k^H_\lambda(0) = \frac{\operatorname{tr}_{x_1}(f_x)^0 \cdot g^0_\lambda - f^0_{2x_2\lambda} \cdot g^0_{x_1}}{f^0_{2x_2k} \cdot g^0_{x_1}}. \tag{4.13}$$

We can apply Theorem 4.5 at the singularly perturbed Hopf points if the nondegeneracy condition (4.4) is satisfied which ensures that the pair of conjugate-complex eigenvalues passes the imaginary axis with nonvanishing velocity. It is not difficult to see that the velocity condition is satisfied if the Bogdanov point of the fast system gives rise to an unfolding diagram as shown in diagram (a) of Figure 1. This diagram shows the standard case.

Concerning the model equation (1.5) (which represents the standard case) the singularly perturbed Hopf points occur at $k^H(\lambda) = -\lambda$, $\lambda < 0$. In particular we

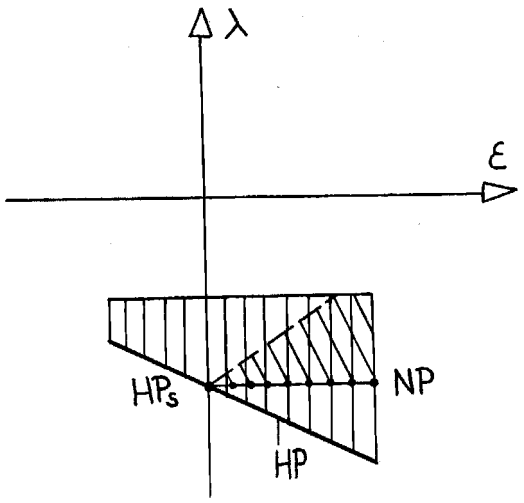


FIG. 7. *The (λ, ϵ)-unfolding diagram near the singularly perturbed Hopf points of the model equation.*

have $k^H(\lambda) \neq 0$ and it is also possible to parametrize the singularly perturbed Hopf points by k. Then Theorem 4.5 can be also applied to $(q, \epsilon) = (\lambda, \epsilon)$-sections for fixed $k > 0$ and we obtain the (λ, ϵ)-unfolding diagram as depicted in Figure 7. The Hopf points HP occur at $\epsilon = 2\lambda(\lambda + k)$, $\lambda < 0$ and the Naimark–Sacker points NP at $(\lambda, \epsilon) = (-k, \epsilon)$, $\epsilon > 0$.

Remember that Figures 6 and 7 show the local (λ, ϵ)-unfoldings of the model equation (1.5) near the singularly perturbed saddle-node and Hopf points. These local diagrams are connected by the parabolic Hopf curve $\epsilon = 2\lambda(\lambda + k)$, $\lambda < 0$ and we obtain for fixed $k > 0$ the global (λ, ϵ)-diagram as indicated in Figure 8. Moreover, the dotted NP-points terminate at the point NP_{-1} that represents a periodic orbit with double Floquet multiplier -1. Hence the two Floquet multipliers different from zero are moving along the unit circle from 1 to -1 during variation along the curve of NP-points (cf. diagram (b) in Figure 3).

Figure 8 shows a detailed (λ, ϵ)-section through the (λ, k, ϵ)-unfolding diagram (a) in Figure 3 for fixed $k > 0$. The homoclinic orbits Ho are not taken into consideration in Figure 8. This will be done in the next section.

5. Homoclinic orbits obtained by a blow-up transformation. In Figure 9 we redraw in detail the numerically calculated surface of homoclinic orbits Ho of the model equation (1.5) from Figure 3(a).

For graphical reasons we changed the direction of the ϵ-axis. Our aim is to verify analytically the hatched upper part of the surface which includes the singularly perturbed homoclinic orbits Ho_s in the $(p_1, p_2) = (\lambda, k)$-plane (dotted curve at $\epsilon = 0$). This is done for the general system (3.7) which was derived from system (2.5) under the assumptions (2.2)–(2.4). In addition we investigate more closely by numerical calculations the lower part of the surface; i.e., we investigate the approach to the border line of open circles where a canard like behavior of the homoclinic orbits seems to occur.

We state the necessary definitions. Consider a system of the form (2.1), but with

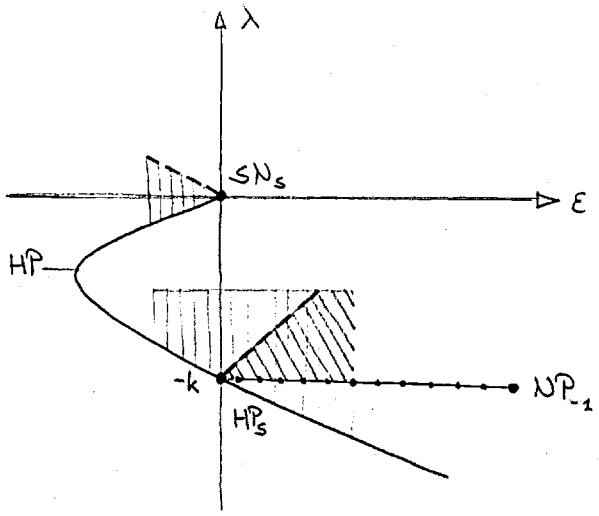


FIG. 8. *The global* (λ, ϵ)*-section through Figure* 3(a) *near the points* SN_s *and* HP_s *for fixed* $k > 0$.

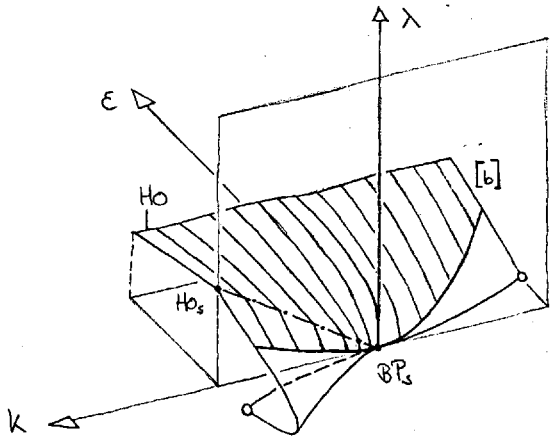


FIG. 9. *The surface of homoclinic orbits for the model equation.*

$u \in \mathbb{R}^n$, $n \geq 2$. Remember that singularly perturbed saddle-node and Hopf points can be also defined for $u \in \mathbb{R}^n$ with the help of center manifold theory (cf. (a) in Figure 4). Here we deal directly with $u \in \mathbb{R}^n$.

Concerning the fast u-system $\dot{u} = F(u, 0, 0)$ at $v = 0$ assume

$$F^0 = 0 \qquad \text{and} \qquad F_u^0 \quad \text{hyperbolic} . \tag{5.1}$$

In addition, assume the existence of a homoclinic orbit $\bar{u}(\tau)$ with base point $u = 0$ such that the only bounded solutions of the homogenous equation

$$\dot{u} = F_u(\bar{u}(\tau), 0, 0, 0) \cdot u \qquad \text{are} \qquad u = c\dot{\bar{u}}, \quad c \in \mathbb{R} . \tag{5.2}$$

Concerning the slow v-perturbation $\dot{v} = \epsilon G(u, v, \epsilon)$ assume as usual the transversality condition (2.4) of the F and G nullclines in the base point $(u, v) = 0$.

DEFINITION 5.1. *A point* $(u,v,\epsilon)=0$ *which satisfies* (5.1), (5.2), *and* (2.4) *is called a singularly perturbed homoclinic base point of the system* (2.1). *The corresponding homoclinic orbit* $(u,v,\epsilon)=(\bar{u}(\tau),0,0)$ *is referred to as singularly perturbed homoclinic orbit.*

Analogous to singularly perturbed saddle-node and Hopf points we investigate singularly perturbed homoclinic orbits within the q-parameter unfolding (4.1). Concerning the corresponding nondegeneracy condition note first that the hyperbolicity of F_u^0 ensures a unique function $(u,v)^s(q,\epsilon)$ such that

$$
\begin{aligned}
&(F,G)((u,v)^s(q,\epsilon),q,\epsilon)=0, \qquad (u,v)^s(0,0)=0,\\
&\begin{pmatrix} u_q^s & u_\epsilon^s \\ v_q^s & v_\epsilon^s \end{pmatrix}(0,0) = -\begin{pmatrix} F_u^0 & F_v^0 \\ G_u^0 & G_v^0 \end{pmatrix}^{-1} \cdot \begin{pmatrix} F_q^0 & F_\epsilon^0 \\ G_q^0 & G_\epsilon^0 \end{pmatrix}.
\end{aligned} \tag{5.3}
$$

DEFINITION 5.2. *The* q*-unfolding* (4.1) *of system* (2.1) *is a generic unfolding of a singularly perturbed homoclinic orbit if*

$$
M := -\int_{-\infty}^{\infty} \Psi^T(\tau)\cdot[F_u,F_v,F_q](\bar{u}(\tau),0,0,0)\cdot \begin{pmatrix} u_q^s(0,0) \\ v_q^s(0,0) \\ 1 \end{pmatrix} d\tau \neq 0, \tag{5.4}
$$

where $\Psi(\tau)$ *denotes the unique bounded solution (up to a constant multiple) of the adjoint equation* $\dot{u}=-F_u^T(\bar{u}(\tau),0,0,0)u$.

The condition (5.4) can be interpreted as a transversality or Melnikov condition along the branch $(u,v)^s(q,0)$ (cf. [23]). We can also interpret it as the regularity condition of an appropriate defining equation in Banach spaces of exponentially weighted functions [2].

Finally, assume for definiteness $(G_u\cdot F_u^{-1}\cdot F_v-G_v)^0<0$ which can always be achieved by time reversal. Then we obtained in [4] the following result with respect to singularly perturbed homoclinic orbits with a generic unfolding.

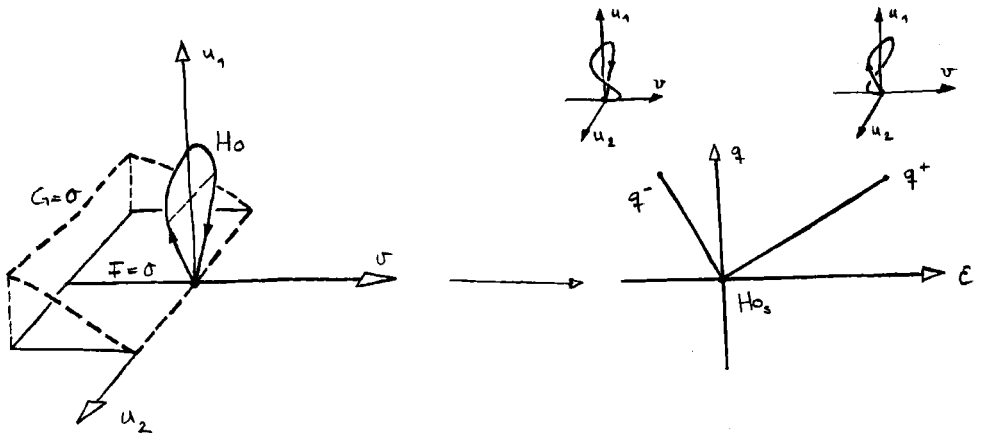
THEOREM 5.3. *There exists a* (q,ϵ)*-unfolding diagram of system* (4.1) *as depicted in diagram* (b) *of Figure* 10.

There exist $\epsilon_0>0$ *and differentiable functions*

$$
\begin{aligned}
q^+(\epsilon) &= q_1^+\cdot\epsilon+O(\epsilon^2), \qquad 0\le\epsilon\le\epsilon_0,\\
q^-(\epsilon) &= q_1^-\cdot\epsilon+O(\epsilon^2), \qquad -\epsilon_0\le\epsilon\le 0
\end{aligned} \tag{5.5}
$$

such that the system (4.1) *has homoclinic orbits* $(u^\pm,v^\pm)(\tau,\epsilon)$ *at* $(q,\epsilon)=(q^\pm(\epsilon),\epsilon)$ *with* $(u^\pm,v^\pm)(\tau,0)=(\bar{u}(\tau),0)$ *and* $q^\pm(0)=0$. *The right and left derivative* q_1^+ *and* q_1^- *can be expressed as*

$$
\begin{aligned}
q_1^+ = \frac{1}{M}\cdot\int_{-\infty}^{\infty}\Psi^T(\tau)\cdot\Bigg\{&-[F_u,F_v](\bar{u}(\tau),0,0,0)\cdot\begin{pmatrix}-(F_u^{-1}F_v)^0\\ 1\end{pmatrix}\cdot\int_\tau^\infty G(\bar{u}(s),0,0,0)ds\\
&+[F_u,F_v,F_\epsilon](\bar{u}(\tau),0,0,0)\cdot\begin{pmatrix}u_\epsilon^s(0,0)\\ v_\epsilon^s(0,0)\\ 1\end{pmatrix}+(F_u^{-1}F_v)^0\cdot G(\bar{u}(\tau),0,0,0)\Bigg\}\,d\tau
\end{aligned} \tag{5.6}
$$



(a) (b)

FIG. 10. (a) *The basic configuration for the fast subsystem* $\dot{u} = F(u, v, 0, 0)$ *and the surface defined by* $G(u, v, 0, 0) = 0$*;* (b) *the* (q, ϵ)*-unfolding diagram of the singularly perturbed homoclinic orbit* Ho_s *and the qualitative shape of the homoclinic orbits that arise for* $\epsilon < 0$ *and* $\epsilon > 0$.

and

$$q_1^- = \frac{1}{M} \cdot \int_{-\infty}^{\infty} \Psi^T(\tau) \cdot \left\{ [F_u, F_v](\bar{u}(\tau), 0, 0, 0) \cdot \begin{pmatrix} -(F_u^{-1} F_v)^0 \\ 1 \end{pmatrix} \cdot \int_{-\infty}^{\tau} G(\bar{u}(s), 0, 0, 0) ds \right.$$

$$\left. + [F_u, F_v, F_\epsilon](\bar{u}(\tau), 0, 0, 0) \cdot \begin{pmatrix} u_\epsilon^s(0,0) \\ v_\epsilon^s(0,0) \\ 1 \end{pmatrix} + (F_u^{-1} F_v)^0 \cdot G(\bar{u}(\tau), 0, 0, 0) \right\} d\tau. \tag{5.7}$$

Concerning the derivatives $(u^s, v^s)_\epsilon(0,0)$ see (5.3). We obtain two nontrivial curves of homoclinic orbits $q^+(\epsilon)$ and $q^-(\epsilon)$ for $\epsilon > 0$ and $\epsilon < 0$, respectively, both of which emanate from $(q, \epsilon) = 0$. Generically, these curves meet transversally at $\epsilon = 0$ due to $q_1^+ \neq q_1^-$. The qualitative shape of the homoclinic orbits along $q^+(\epsilon)$ and $q^-(\epsilon)$ is also indicated in diagram (b).

As usual we indicate in diagram (a) of Figure 10 the basic configuration for the fast subsystem $\dot{u} = F(u, v, 0, 0)$, $\dot{v} = 0$. In contrast to singularly perturbed saddle-node and Hopf points we do not require a velocity condition with respect to v (cf. (2.3) and (4.4)); i.e., we do not assume that the stable and unstable manifold pass each other in the homoclinic orbit Ho with nonvanishing velocity along the branch of stationary points $F(u, v, 0, 0) = 0$. The branch of stationary points (parametrized by v) of the system $\dot{u} = F(u, v, 0, 0)$ is ensured by the condition that F_u^0 is hyperbolic.

The fact that we can omit the velocity condition allows us to treat the singularly perturbed homoclinic orbits of the FitzHugh–Nagumo equation (1.3) which occur again at $c = 0$ within the family of Hamiltonian systems $\dot{u}_1 = u_2$, $\dot{u}_2 = u_1(u_1 - a)(u_1 - 1) + v$, but for now $a > 0$. In that special case every stationary point in diagram (a) of Figure 10 is a homoclinic base point of the fast u-system and we obtain a continuum of homoclinic orbits. Theorem 5.3 shows us that this degeneracy represents no special difficulties. From the continuum of homoclinic orbits only the homoclinic orbit whose base point satisfies $G = 0$ is continued to $\epsilon \neq 0$. This orbit

just represents the singularly perturbed homoclinic orbit within the continuum of homoclinic orbits. The remaining homoclinic orbits are destroyed.

Similar results as stated in Theorem 5.3 appear in [28], [18] where general transversality arguments are used to prove the pure existence of homoclinic orbits for $\epsilon \neq 0$. The precise quantitative behavior of the homoclinic solutions, for example the ϵ-expansions or the asymptotic time estimates of $(u^{\pm}, v^{\pm})(\tau, \epsilon)$ can be found in [4].

We return to the investigation of singularly perturbed Bogdanov points. In the last section we determined the unfolding diagrams of singularly perturbed saddle-node and Hopf points. We obtained some information on 2-dimensional sections through the 3-dimensional unfolding space of a singularly perturbed Bogdanov point. We could also follow this way when dealing with singularly perturbed homoclinic orbits. However, we get additional insight if we start from the center of the unfolding diagram, i.e., from the singularly perturbed Bogdanov point at $(p_1, p_2, \epsilon) = (\lambda, k, \epsilon) = 0$. This is done by an appropriate blow-up.

5.1. The blow-up. We start again from the pretransformed (x, y)-system (3.7). The blow-up transformation between the old variables $(x_1, x_2, y, \epsilon, \lambda, k, t)$ and the new variables $(u_1, u_2, v, \bar{\epsilon}, \bar{\lambda}, \bar{k}, \tau)$ reads

$$
\begin{array}{lll}
\text{coordinates :} & (x_1\ ,\ x_2\ ,\ y) = (c_1 \cdot \bar{\lambda}^2 \cdot u_1\ ,\ c_2 \cdot \bar{\lambda}^3 \cdot u_2\ ,\ c_3 \cdot \bar{\lambda}^4 \cdot v), \\
\text{parameters :} & (\epsilon\ ,\ \ \lambda\ ,\ \ k) = (c_4 \cdot \bar{\lambda}^3 \cdot \bar{\epsilon}\ ,\ \ c_5 \cdot \bar{\lambda}^2\ ,\ \ \ \ c_6 \cdot \bar{\lambda}^2 \cdot \bar{k}), \\
\text{time :} & t \qquad\qquad\quad = c_7 \cdot \bar{\lambda}^{-1} \cdot \tau,
\end{array}
\tag{5.8}
$$

where we used u and v again to denote the new coordinates. The constants $c_1, \ldots, c_7$ will be chosen later on. Essentially this $\bar{\lambda}$-scaling is a direct extension of the well-known scaling near a regularly perturbed Bogdanov point [15]. However, note that we maintain a linear dependence between the variables y and v. This seems appropriate because v represents a dynamic variable for $\bar{\epsilon} \neq 0$ and it is convenient to decouple the parameter transformation between (ϵ, λ, k) and $(\bar{\epsilon}, \bar{\lambda}, \bar{k})$ from the coordinate transformation between (x_1, x_2, y) and (u_1, u_2, v). Then a straightforward calculation yields the transformed system

$$
\begin{aligned}
\dot{u}_1 &= u_2 + \bar{\lambda} \cdot (a_1 u_1^2 + a_2 v) + O(\bar{\lambda}^2) \\
&=: F_1(u_1, u_2, v, \bar{\lambda}, \bar{k}, \bar{\epsilon}), \\
\dot{u}_2 &= u_1^2 - v + \bar{\lambda} \cdot (b_1 u_1 u_2 + b_2 \bar{k} u_2 + b_3 u_2) + O(\bar{\lambda}^2) \\
&=: F_2(u_1, u_2, v, \bar{\lambda}, \bar{k}, \bar{\epsilon}), \\
\dot{v} &= \bar{\epsilon} \cdot [u_1 - 2 + O(\bar{\lambda})] \\
&=: \bar{\epsilon} \cdot G(u_1, u_2, v, \bar{\lambda}, \bar{k}, \bar{\epsilon})
\end{aligned}
\tag{5.9}
$$

which maintains the ordinary singular perturbation form. The constants in (5.8) are chosen according to

$$
\begin{array}{lllll}
\text{coordinates :} & c_1 = \frac{1}{2}\ f^0_{2x_1x_1}, & c_2 = c_1^2, & c_3 = -c_1^3/f^0_{2y}, \\
\text{parameters :} & c_4 = -c_1^3/(f^0_{2y} \cdot g^0_{x_1}), & c_5 = -2\ c_1\ g^0_{x_1}/g^0_{\lambda}, & c_6 = c_1, \\
\text{time :} & c_7 = 1/c_1.
\end{array}
\tag{5.10}
$$

This yields the following coefficients for the linear $\bar\lambda$-terms in (5.9):

$$\begin{aligned} &a_1 = \tfrac{1}{2}\, f^0_{1x_1x_1}, \quad a_2 = -\tfrac{1}{2}\, f^0_{2x_1x_1}\cdot f^0_{1y}/f^0_{2y}, \\ &b_1 = f^0_{2x_1x_2}, \qquad b_2 = f^0_{2x_2k}, \qquad\qquad b_3 = -2\, f^0_{2x_2\lambda}\cdot g^0_{x_1}/g^0_\lambda. \end{aligned} \tag{5.11}$$

According to (3.3) and (3.8), (3.10) these coefficients are completely expressed in terms of derivatives of the original system (2.5) evaluated at the singularly perturbed Bogdanov point. Under the assumptions (2.2)–(2.4) and (3.1) the coefficients may assume arbitrary values. In the next step we shall impose conditions on the coefficients which ensure the existence of homoclinic orbits of the system (5.9) for $\bar\epsilon \neq 0$. In particular the difference between the assumption (3.1) and the stronger assumption (2.7) of a generic unfolding is analyzed.

First we are looking for homoclinic orbits of the fast u-system $\dot u = F(u, v, \bar\lambda, \bar k, 0)$ at $\bar\epsilon = 0$. Second we identify among these orbits the homoclinic orbits whose base points satisfy $G = 0$; i.e., we are looking for singularly perturbed homoclinic orbits of the blown-up system (5.9). Third we perform the actual singular perturbation; i.e., we continue the singularly perturbed homoclinic orbits to $\bar\epsilon \neq 0$. This is done with the help of the Melnikov condition (5.4) which represents the nondegeneracy condition of the unfolding.

5.2. Homoclinic orbits at $\bar\epsilon = 0$. At $\bar\epsilon = 0$, $\bar\lambda = 0$, and $\bar k \in \mathbb{R}$, the fast u-system $\dot u_1 = u_2,\ \dot u_2 = u_1^2 - v$ in (5.9) represents a well-known v-dependent family of Hamiltonian systems [15] with corresponding family of homoclinic orbits

$$\begin{aligned} \bar u_1(\tau, v) &= v^{\frac12}\cdot\left[1 - 3\ \mathrm{sech}^2\left(2^{-\frac12} v^{\frac14}\tau\right)\right], \\ \bar u_2(\tau, v) &= 3\sqrt{2} v^{\frac34}\cdot \mathrm{sech}^2\left(2^{-\frac12} v^{\frac14}\tau\right)\cdot \tanh\left(2^{-\frac12} v^{\frac14}\tau\right) \end{aligned} \tag{5.12}$$

for $v > 0$. Note that the base points $\bar u(\infty, v)$ of the homoclinic orbits are

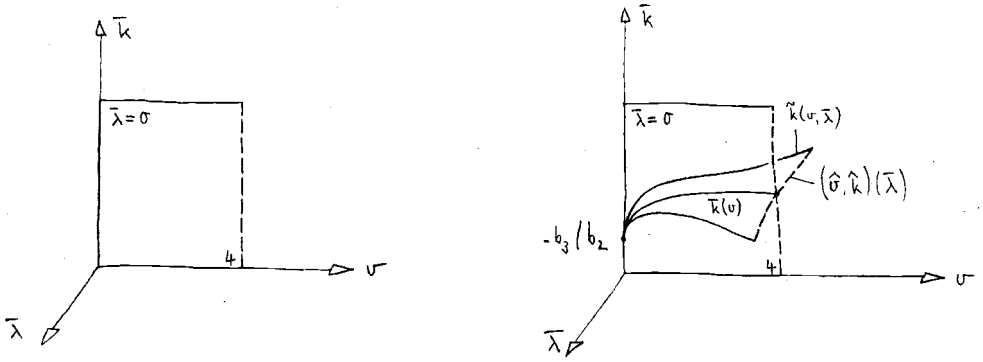
$$\bar u_1(\infty, v) = \sqrt{v}, \qquad \bar u_2(\infty, v) = 0, \qquad v > 0. \tag{5.13}$$

Hence in the 3-dimensional $(v, \bar\lambda, \bar k)$-parameter space of the u-system $\dot u = F(u, v, \bar\lambda, \bar k, 0)$ we obtain a surface of homoclinic orbits at $\bar\lambda = 0$ as indicated in diagram (a) of Figure 11. Obviously this trivial surface exists for arbitrary values of the coefficient b_2.

Next we continue some of these homoclinic orbits to $\bar\lambda \neq 0$. This leads to diagram (b) in Figure 11 under the additional restriction $b_2 = f^0_{2x_2k} \neq 0$ which is satisfied for a generic unfolding of a singularly perturbed Bogdanov point (cf. (3.10)). Without the extra v-parameter, bifurcation results of this type are derived in [7], [15], and [3]; therefore we restrict ourselves to the special features of our situation.

First the bifurcation condition which is equivalent to a vanishing Melnikov function with respect to $\bar\lambda$ reads

$$\begin{aligned} 0 &= -\int_{-\infty}^{\infty} [-\dot{\bar u}_2,\ \dot{\bar u}_1]\cdot F_{\bar\lambda}(\bar u, v, 0, \bar k, 0)\ d\tau \\ &= -\int_{-\infty}^{\infty} -\ddot{\bar u}_1\cdot(a_1\bar u_1^2 + a_2 v) + \dot{\bar u}_1^2\cdot(b_1\bar u_1 + b_2\bar k + b_3)\ d\tau \\ &= -(2a_1 + b_1)\cdot\int_{-\infty}^{\infty} \dot{\bar u}_1^2\bar u_1\ d\tau - (b_2\bar k + b_3)\cdot\int_{-\infty}^{\infty}\dot{\bar u}_1^2\ d\tau \\ &= \sqrt{2}\tfrac{24}{35} v^{\frac54}\cdot[5(2a_1 + b_1)\sqrt{v} - 7(b_2\bar k + b_3)] =: M(v, \bar k). \end{aligned} \tag{5.14}$$



$b_2\in\mathbb{R}$ (a) $b_2\neq 0$ (b)

FIG. 11. *The system* (5.9) *at* $\bar{\epsilon}=0$: (a) *The trivial surface of homoclinic orbits obtained by the blow-up;* (b) *the nontrivial surface of homoclinic orbits in case of a generic unfolding.*

In the case of $b_2 \neq 0$ we obtain $M(v,\bar{k})=0$, $v>0$ iff

$$\bar{k} = \frac{5(2a_1+b_1)\sqrt{v}-7b_3}{7b_2} =: \bar{k}(v). \tag{5.15}$$

Finally the nondegeneracy condition for bifurcations along $\bar{k}(v)$ reads

$$0 \neq \int_{-\infty}^{\infty} [-\dot{\bar{u}}_2,\ \dot{\bar{u}}_1]\cdot F_{\bar{\lambda}\bar{k}}(\bar{u},v,0,\bar{k}(v),0)\ d\tau = b_2\cdot\int_{-\infty}^{\infty}\dot{\bar{u}}_1^2\ d\tau \tag{5.16}$$

which is true according to $b_2\neq 0$. Hence we obtain a nontrivial surface of homoclinic orbits

$$[u,v,\bar{\lambda},\bar{k}] = [\tilde{u}(\tau,v,\bar{\lambda}),v,\bar{\lambda},\tilde{k}(v,\bar{\lambda})], \qquad v>0, \qquad |\bar{\lambda}|\ll 1 \tag{5.17}$$

with $\tilde{u}(\tau,v,0)=\bar{u}(\tau,v)$ and $\tilde{k}(v,0)=\bar{k}(v)$ (cf. diagram (b) in Figure 11). Summarizing, the fast u-system $\dot{u}=F(u,v,\bar{\lambda},\bar{k},0)$ has the trivial surface (5.12) of homoclinic orbits at $\bar{\lambda}=0$ which is intersected at $(v,\bar{\lambda},\bar{k})=(v,0,\bar{k}(v))$ by the nontrivial surface (5.17) in case of $b_2\neq 0$.

Up to now we performed a well-known bifurcation analysis of homoclinic orbits. In the next section we extend this bifurcation analysis to the full (x,y)-system (5.9) with respect to the singular perturbation parameter $\bar{\epsilon}$; i.e., we determine in Figure 11 the homoclinic orbits that can be continued to $\bar{\epsilon}\neq 0$.

A generic unfolding of a singularly perturbed Bogdanov point implies $b_2 = f^0_{2x_2k}\neq 0$ and diagram (b) in Figure 11 occurs. However, our approach allows us to treat the degenerate case $b_2=0$ almost without extra effort; therefore we include this case.

5.3. The bifurcation analysis with respect to $\bar{\epsilon}$. Our aim is to apply Theorem 5.3 to the blown-up system (5.9). This is done twice with respect to $(q,\epsilon)=(\bar{\lambda},\bar{\epsilon})$ and $(q,\epsilon)=(\bar{k},\bar{\epsilon})$, respectively. In section 5.2 we determined the homoclinic orbits at $\bar{\epsilon}=0$. Next we identify the homoclinic orbits whose base points satisfy $G(u,v,\bar{\lambda},\bar{k},0)=0$ in addition. This simple condition represents the bifurcation equation.

Recalling that $(u_1,u_2,v,\bar{\lambda},\bar{k})=(\sqrt{v},0,v,0,\bar{k})$ are homoclinic base points of the trivial surface (cf. (5.13)) we obtain

$$G[\sqrt{v},0,v,0,\bar{k},0]=\sqrt{v}-2=0 \qquad \text{iff} \qquad v=4, \qquad \bar{k}\in\mathbb{R}. \tag{5.18}$$

This leads to the vertical broken lines in the diagrams of Figure 11 at $v = 4$, $\bar\lambda = 0$. Concerning the nontrivial homoclinic orbits in the case of $b_2 \neq 0$ we consider (cf. (5.17))

$$\begin{aligned} G[\tilde u(\infty, v, \bar\lambda), v, \bar\lambda, \tilde k(v, \bar\lambda), 0] &= \tilde u_1(\infty, v, 0) - 2 + O(\bar\lambda) \\ &= \sqrt{v} - 2 + O(\bar\lambda) = 0 \end{aligned} \tag{5.19}$$

and obtain a locally unique function $\hat v(\bar\lambda)$ with $\hat v(0) = 4$ satisfying

$$G[\tilde u(\infty, \hat v(\bar\lambda), \bar\lambda), \hat v(\bar\lambda), \bar\lambda, \tilde k(\hat v(\bar\lambda), \bar\lambda), 0] = 0. \tag{5.20}$$

We define $(\hat u, \hat k)(\bar\lambda) := [\tilde u(\infty, \hat v(\bar\lambda), \bar\lambda), \tilde k(\hat v(\bar\lambda), \bar\lambda)]$ and denote the corresponding homoclinic orbits by $\hat u(\tau, \bar\lambda)$.

Summarizing, broken lines in Figure 11 indicate the homoclinic orbits at $\bar\epsilon = 0$ which satisfy the bifurcation equation $G = 0$ and it remains to verify the nondegeneracy conditions in (5.1), (5.2), (2.4), and (5.4) for proving bifurcation with respect to $\bar\epsilon$ with the help of Theorem 5.3.

First we obtain from (5.9) the regularity of the matrix

$$\begin{pmatrix} F_u & F_v \\ G_u & G_v \end{pmatrix} (\sqrt{v}, 0, v, 0, \bar k, 0) = \begin{pmatrix} 0 & 1 & 0 \\ 2\sqrt{v} & 0 & -1 \\ 1 & 0 & 0 \end{pmatrix}$$

as desired (cf. (2.4)). Thus we have a unique surface of steady states $(u, v)^s(\bar\lambda, \bar k, \bar\epsilon)$ with

$$\begin{aligned} &(F, G)[(u, v)^s(\bar\lambda, \bar k, \bar\epsilon), \bar\lambda, \bar k, \bar\epsilon] = 0 \qquad \text{and} \\ &(u, v)^s(0, \bar k, 0) \equiv (2, 0, 4), \qquad (u, v)^s(\bar\lambda, \hat k(\bar\lambda), 0) = (\hat u, \hat v)(\bar\lambda) \end{aligned} \tag{5.21}$$

(cf. (5.3)). Second, the condition (5.2) is satisfied along the broken lines in Figure 11 at $\bar\lambda = 0$ as can be seen from the behavior of the Wronski-determinant of the system $\dot u = F_u(\bar u(\tau, 4), 4, 0, \bar k, 0)u$ as τ tends to infinity. Concerning the nontrivial homoclinic orbits in Figure 11 this property follows locally from a perturbation argument.

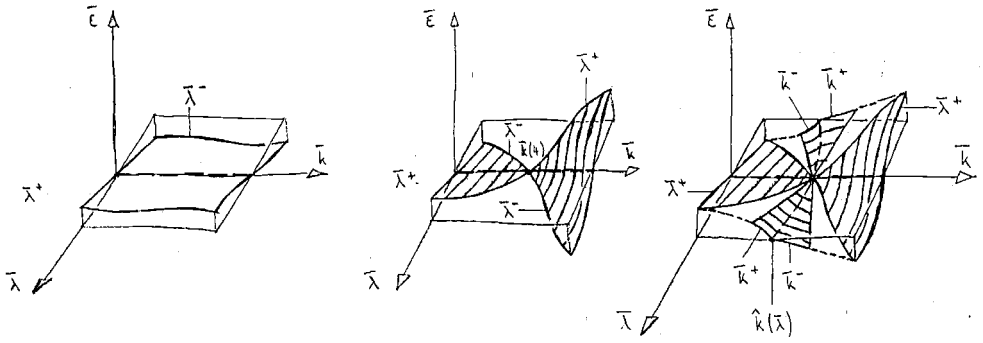
Finally, we verify the Melnikov condition (5.4) separately along $[u, v, \bar\lambda, \bar k, \bar\epsilon] = [\bar u(\tau, 4), 4, 0, \bar k, 0]$ with respect to $(q, \epsilon) = (\bar\lambda, \bar\epsilon)$ and along $[u, v, \bar\lambda, \bar k, \bar\epsilon] = [\hat u(\tau, \bar\lambda), \hat v(\bar\lambda), \bar\lambda, \hat k(\bar\lambda), 0]$ with respect to $(q, \epsilon) = (\bar k, \bar\epsilon)$.

First we continue the homoclinic orbits $[u, v, \bar\lambda, \bar k, \bar\epsilon] = [\bar u(\tau, 4), 4, 0, \bar k, 0]$ in the $(\bar\lambda, \bar\epsilon)$-plane. In this case the Melnikov condition reads

$$\begin{aligned} 0 \neq &- \int_{-\infty}^{\infty} [-\dot{\bar u}_2,\ \dot{\bar u}_1](\tau, 4) \cdot [F_u, F_v, F_{\bar\lambda}](\bar u(\tau, 4), 4, 0, \bar k, 0) \cdot \begin{pmatrix} u^s_{\bar\lambda}(0, \bar k, 0) \\ v^s_{\bar\lambda}(0, \bar k, 0) \\ 1 \end{pmatrix} d\tau \\ &= - \int_{-\infty}^{\infty} [-\dot{\bar u}_2,\ \dot{\bar u}_1](\tau, 4) \cdot F_{\bar\lambda}(\bar u(\tau, 4), 4, 0, \bar k, 0)\ d\tau = M(4, \bar k) \end{aligned} \tag{5.22}$$

(cf. (5.14)). Note that $\int_{-\infty}^{\infty} \dot{\bar u}_1\ d\tau = \int_{-\infty}^{\infty} \dot{\bar u}_2\ d\tau = \int_{-\infty}^{\infty} \dot{\bar u}_1 \bar u_1\ d\tau = 0$. According to (5.14), (5.15) this nondegeneracy condition is satisfied for

$$\bar k \in \mathbb{R} \qquad \text{in the case of} \qquad b_2 = 0, \qquad 10(2a_1 + b_1) - 7b_3 =: \gamma \neq 0 \tag{5.23}$$



(a) (b) (c)

FIG. 12. *The surfaces of homoclinic orbits of the blown-up system* (5.9): (a) $b_2 = 0$, $\gamma \neq 0$, (b) *and* (c) $b_2 \neq 0$.

and for

$$\bar{k} \neq \bar{k}(4) = \frac{\gamma}{7b_2} \qquad \text{in the case of} \qquad b_2 \neq 0. \tag{5.24}$$

Thus, we can apply Theorem 5.3. In the $(\bar{\lambda}, \bar{k}, \bar{\epsilon})$-parameter space of the system (5.9) we obtain smooth $\bar{k}$-dependent surfaces

$$\bar{\lambda} = \bar{\lambda}^+(\bar{k}, \bar{\epsilon}), \quad 0 \leq \bar{\epsilon} \leq \bar{\epsilon}_0(\bar{k}) \qquad \text{and} \qquad \bar{\lambda} = \bar{\lambda}^-(\bar{k}, \bar{\epsilon}), \quad -\bar{\epsilon}_0(\bar{k}) \leq \bar{\epsilon} \leq 0 \tag{5.25}$$

of homoclinic orbits which branch off from the $\bar{k}$-axis for $\bar{k} \in \mathbb{R}$ in the case of $b_2 = 0$, $\gamma \neq 0$ (cf. diagram (a) in Figure 12), and for $\bar{k} \in \mathbb{R} \setminus \{\bar{k}(4)\}$ in the case of $b_2 \neq 0$ (cf. diagram (b) in Figure 12). It should be noted that the proof in [4] does not describe in detail the limiting behavior of $\bar{\epsilon}_0(\bar{k})$ as $\bar{k}$ tends to $\pm\infty$ or as $\bar{k}$ tends to $\bar{k}(4)$ in the case of $b_2 \neq 0$. We only know $\bar{\epsilon}_0(\bar{k}) \to 0$ in these situations.

Further it is not difficult to see the antisymmetry $\bar{\lambda}^+(\bar{k}, \bar{\epsilon}) = -\bar{\lambda}^-(\bar{k}, -\bar{\epsilon})$ which results from a certain symmetry of the scaling transformation (5.8) and a uniqueness argument (cf. [3]). In particular the surfaces are smooth along the $\bar{k}$-axis, i.e., the derivatives at $\bar{\epsilon} = 0$ satisfy $\bar{\lambda}_1^-(\bar{k}) = \bar{\lambda}_1^+(\bar{k})$ (cf. (5.5)) and the homoclinic orbits which correspond to $\bar{\lambda} = \bar{\lambda}^+(\bar{k}, \bar{\epsilon})$ and $\bar{\lambda} = -\bar{\lambda}^-(\bar{k}, -\bar{\epsilon})$ represent the same homoclinic orbit of the unscaled system (3.7). Hence we can restrict ourselves to

$$\bar{\lambda}^+(\bar{k}, \bar{\epsilon}) = \bar{\lambda}_1^+(\bar{k}) \cdot \bar{\epsilon} + O(\bar{\epsilon}^2), \quad 0 \leq \bar{\epsilon} \leq \bar{\epsilon}_0(\bar{k}). \tag{5.26}$$

The derivative $\bar{\lambda}_1^+(\bar{k})$ at $\bar{\epsilon} = 0$ follows after some calculations from (5.6) according to

$$\bar{\lambda}_1^+(\bar{k}) = \frac{1}{M(4, \bar{k})} \cdot \int_{-\infty}^{\infty} [\bar{u}_1(\tau, 4) - 2]^2 \, d\tau = \frac{35}{4} \cdot \frac{1}{\gamma - 7b_2\bar{k}}. \tag{5.27}$$

In the case of $b_2 \neq 0$ this derivative has a pole at $\bar{k} = \bar{k}(4)$ and converges to zero as $\bar{k} \to \pm\infty$. In the case of $b_2 = 0$ we obtain the identity $\bar{\lambda}_1^+(\bar{k}) \equiv 35/(4\gamma)$.

It remains to verify in the case of $b_2 \neq 0$ the Melnikov condition (5.4) along the homoclinic orbits $[u, v, \bar{\lambda}, \bar{k}, \bar{\epsilon}] = [\hat{u}(\tau, \bar{\lambda}), \hat{v}(\bar{\lambda}), \bar{\lambda}, \hat{k}(\bar{\lambda}), 0]$ with respect to $(q, \epsilon) = (\bar{k}, \bar{\epsilon})$. Let $\psi(\tau, \bar{\lambda})$ denote the unique bounded solution (up to a scalar multiple) of the corresponding adjoint equation $\dot{u} = -F_u^T[\hat{u}(\tau, \bar{\lambda}), \hat{v}(\bar{\lambda}), \bar{\lambda}, \hat{k}(\bar{\lambda}), 0]u$. $\psi(\tau, \bar{\lambda})$ can be

chosen to vary smoothly in $(\tau, \bar{\lambda})$ and to satisfy $\psi^T(\tau, 0) = [-\dot{\hat{u}}_2, \dot{\hat{u}}_1](\tau, 4)$. Then the nondegeneracy condition reads

$$0 \neq -\int_{-\infty}^{\infty} \psi^T(\tau, \bar{\lambda}) \cdot [F_u, F_v, F_{\bar{k}}][\hat{u}(\tau, \bar{\lambda}), \hat{v}(\bar{\lambda}), \bar{\lambda}, \hat{k}(\bar{\lambda}), 0] \cdot \begin{pmatrix} u_{\bar{k}}^s(\bar{\lambda}, \hat{k}(\bar{\lambda}), 0) \\ v_{\bar{k}}^s(\bar{\lambda}, \hat{k}(\bar{\lambda}), 0) \\ 1 \end{pmatrix} d\tau \tag{5.28}$$

$$=: \hat{M}(\bar{\lambda}).$$

At $\bar{\lambda} = 0$ we have the identities $\hat{M}(0) = 0$ due to $(u_{\bar{k}}^s, v_{\bar{k}}^s)(0, 0, 0) = 0$ (cf. (5.21)) and $F_{\bar{k}}[\hat{u}(\tau, 0), 4, 0, \hat{k}(0), 0] = 0$ due to (5.9). Further, differentiation at $\bar{\lambda} = 0$ shows after some calculations

$$\begin{aligned} \hat{M}'(0) &= -\int_{-\infty}^{\infty} \psi^T(\tau, 0) \cdot \frac{d}{d\bar{\lambda}} F_{\bar{k}}[\hat{u}(\tau, \bar{\lambda}), \hat{v}(\bar{\lambda}), \bar{\lambda}, \hat{k}(\bar{\lambda}), 0]_{|\bar{\lambda}=0} \, d\tau \\ &= -b_2 \cdot \int_{-\infty}^{\infty} \dot{\hat{u}}_1^2(\tau, 4) \, d\tau \neq 0 \end{aligned} \tag{5.29}$$

and we can apply again Theorem 5.3. According to (5.5) this yields the surfaces

$$\bar{k} = \bar{k}^+(\bar{\lambda}, \bar{\epsilon}), \quad 0 \leq \bar{\epsilon} \leq \bar{\epsilon}_0(\bar{\lambda}) \qquad \text{and} \qquad \bar{k} = \bar{k}^-(\bar{\lambda}, \bar{\epsilon}), \quad -\bar{\epsilon}_0(\bar{\lambda}) \leq \bar{\epsilon} \leq 0 \tag{5.30}$$

of homoclinic orbits which branch off from the curve $[u, v, \bar{\lambda}, \bar{k}, \bar{\epsilon}] = [\hat{u}(\tau, \bar{\lambda}), \hat{v}(\bar{\lambda}), \bar{\lambda}, \hat{k}(\bar{\lambda}), 0]$ for $0 < |\bar{\lambda}| \ll 1$. Hence we can extend diagram (b) in Figure 12 as depicted in diagram (c). Note that in general we cannot expect $\bar{k}^+(\bar{\lambda}, \bar{\epsilon})$ and $\bar{k}^-(\bar{\lambda}, \bar{\epsilon})$ to form a smooth surface along $\hat{k}(\bar{\lambda})$. The precise behavior of $\bar{\epsilon}_0(\bar{\lambda})$ remains unknown as $\bar{\lambda}$ tends to zero. The surfaces show again a symmetry given by $\bar{k}^+(\bar{\lambda}, \bar{\epsilon}) = \bar{k}^-(-\bar{\lambda}, -\bar{\epsilon})$ and we can restrict to $\bar{k}^+(\bar{\lambda}, \bar{\epsilon})$.

The different surfaces $\bar{\lambda}^{\pm}(\bar{k}, \bar{\epsilon})$ and $\bar{k}^{\pm}(\bar{\lambda}, \bar{\epsilon})$ of homoclinic orbits in diagram (c) of Figure 12 are presumably parts of a common surface of homoclinic orbits as indicated by broken lines. Numerical calculations strongly confirm this conjecture.

For closing analytically the gaps between the different surfaces it would be necessary to generalize the bifurcation result in [4] which concentrates on singularly perturbed homoclinic orbits with a generic unfolding. Remember that points of this type occur in Figure 12 at $\bar{\epsilon} = 0$ along the $\bar{k}$-axis for $\bar{k} \neq \bar{k}(4)$ and along $\hat{k}(\bar{\lambda})$ for $\bar{\lambda} \neq 0$. The singularly perturbed homoclinic orbit at the bifurcation point of the two curves is not generically unfolded and the theory is not applicable.

5.4. The back transformation. In this section we transform the results obtained for the blown-up system (5.9) back to the system (3.7); i.e., we interpret the surfaces from Figure 12 within the (λ, k, ϵ)-unfolding space of our singularly perturbed Bogdanov point using the scaling transformation (5.8). Essentially this yields the hatched upper part of the surface of homoclinic orbits in Figure 9.

First we indicate the transformation of $\bar{\lambda}^+(\bar{k}, \bar{\epsilon})$ in the case of $b_2 \neq 0$. This inequality occurs at a singularly perturbed Bogdanov point with a generic unfolding. According to (5.25)–(5.27) we obtain

$$\bar{\lambda}^+(\bar{k}, \bar{\epsilon}) = \frac{35}{4} \cdot \frac{1}{\gamma - 7b_2\bar{k}} \cdot \bar{\epsilon} + O(\bar{\epsilon}^2), \qquad \bar{\epsilon} \in [0, \bar{\epsilon}_0(\bar{k})], \qquad \bar{k} \neq \bar{k}(4) \tag{5.31}$$

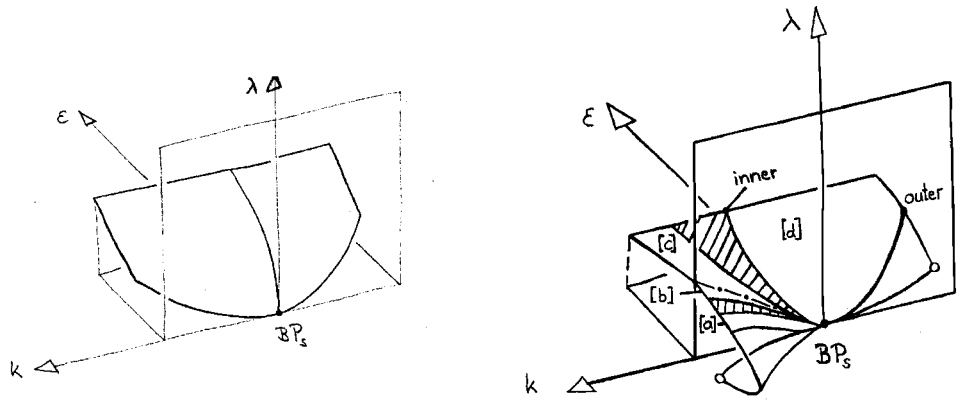


FIG. 13. *The different surfaces* [a]–[d] *(see text) of homoclinic orbits of the pretransformed system* (3.7)*;* (a) $b_2 = 0$, $\gamma \neq 0$*;* (b) $b_2 \neq 0$.

or equivalently

$$\bar{\epsilon}^+(\bar{k}, \bar{\lambda}) = \tfrac{4}{35}(\gamma - 7b_2\bar{k}) \cdot \bar{\lambda} + O(\bar{\lambda}^2), \qquad \bar{k} \neq \bar{k}(4), \tag{5.32}$$

where $\bar{\epsilon}^+(\bar{k}, \cdot)$ denotes the inverse function of $\bar{\lambda}^+(\bar{k}, \cdot)$ and $\bar{\lambda}$ is chosen according to

$$\bar{\lambda} \in \begin{cases} [0,\ \bar{\lambda}_0(\bar{k})] & \text{in the case of} \quad \gamma - 7b_2\bar{k} > 0, \\ [-\bar{\lambda}_0(\bar{k}),\ 0] & \text{in the case of} \quad \gamma - 7b_2\bar{k} < 0 \end{cases} \tag{5.33}$$

with $0 < \bar{\lambda}_0(\bar{k}) \ll 1$ and $\bar{\lambda}_0(\bar{k}) \to 0$ as $\bar{k} \to \pm\infty$ or as $\bar{k} \to \bar{k}(4)$. Then recalling the scaling transformation (5.8), the system (3.7) has homoclinic orbits at parameter values

$$(\epsilon, \lambda, k) = (\tfrac{4}{35}c_4(\gamma - 7b_2\bar{k}) \cdot \bar{\lambda}^4 + O(\bar{\lambda}^5), \quad c_5 \cdot \bar{\lambda}^2, \quad c_6\bar{k} \cdot \bar{\lambda}^2) \tag{5.34}$$

with $\bar{k} \in \mathbb{R} \setminus \{\bar{k}(4)\}$ and $\bar{\lambda}$ chosen according to (5.33). For fixed $\bar{k}$ we obtain a curve in the (λ, k, ϵ)-space parametrized by $\bar{\lambda}$. The projection of this curve onto the (λ, k)-plane forms the straight line

$$k = \frac{c_6}{c_5}\bar{k} \cdot \lambda, \qquad \lambda \in \begin{cases} [0, c_5\bar{\lambda}_0^2(\bar{k})] & \text{in the case of} \quad c_5 > 0 \\ [c_5\bar{\lambda}_0^2(\bar{k}), 0] & \text{in the case of} \quad c_5 < 0 \end{cases}. \tag{5.35}$$

In contrast to this the projection of the curve onto the (λ, ϵ)-plane yields the parabola

$$\epsilon = \frac{4}{35}\frac{c_4}{c_5^2}(\gamma - 7b_2\bar{k}) \cdot \lambda^2 + O(\lambda^2\sqrt{\lambda}). \tag{5.36}$$

During variation of $\bar{k} \in (-\infty, \bar{k}(4))$ and $\bar{k} \in (\bar{k}(4), \infty)$ we obtain the surfaces [a] and [d] as indicated in diagram (b) of Figure 13. These two surfaces represent parts of the hatched upper part in Figure 9. Remember that Figures 9 and 13 show the (λ, k, ϵ)-unfolding space of a singularly perturbed Bogdanov point with a generic unfolding.

The outer and inner border lines of the surfaces [a] and [d] show the limiting behavior

$$\frac{\lambda}{k} = \frac{c_5\bar{\lambda}_0^2(\bar{k})}{c_6\bar{\lambda}_0^2(\bar{k})\cdot\bar{k}} = \frac{c_5}{c_6}\cdot\frac{1}{\bar{k}} \to \begin{cases} 0 & \text{as} \quad \bar{k}\to\pm\infty, \\ \dfrac{c_5}{c_6}\cdot\dfrac{1}{\bar{k}(4)} & \text{as} \quad \bar{k}\to\bar{k}(4) \end{cases}. \tag{5.37}$$

In particular we have $\lambda/k = o(1)$ if we approach the origin along the outer border lines. In the case of $\bar{k}(4) = 0$ the inner border lines are tangentially touching the λ-axis. This situation occurs in the FitzHugh–Nagumo equation that is treated in detail in [4].

Next we consider $\bar{\lambda}^+(\bar{k},\bar{\epsilon})$ in the case of $b_2 = 0$, $\gamma \neq 0$ (cf. diagram (a) in Figure 12). In this situation we can choose $\bar{k}\in\mathbb{R}$ and obtain a unique surface in the (λ,k,ϵ)-parameter space as indicated in diagram (a) of Figure 13. More precisely there exist homoclinic orbits at parameter values

$$(\epsilon,\ \lambda,\ k) = (\tfrac{4}{35}c_4\gamma\cdot\bar{\lambda}^4 + O(\bar{\lambda}^5),\quad c_5\cdot\bar{\lambda}^2,\quad c_6\bar{k}\cdot\bar{\lambda}^2) \tag{5.38}$$

with $\bar{k}\in\mathbb{R}$ and $\bar{\lambda}$ chosen according to

$$\bar{\lambda} \in \begin{cases} [0,\ \bar{\lambda}_0(\bar{k})] & \text{in the case of} \quad \gamma > 0, \\ [-\bar{\lambda}_0(\bar{k}),\ 0] & \text{in the case of} \quad \gamma < 0 \end{cases}$$

with $\bar{\lambda}_0(\bar{k}) \to 0$ as $\bar{k}\to\pm\infty$.

In principle the back transformation of the two surfaces $\bar{k}^+(\bar{\lambda},\bar{\epsilon})$, $0 < \bar{\lambda} \ll 1$ and $\bar{k}^+(\bar{\lambda},\bar{\epsilon})$, $-1 \ll \bar{\lambda} < 0$ from diagram (c) in Figure 12 works along the same lines as the back transformation of $\bar{\lambda}^+(\bar{k},\bar{\epsilon})$; therefore we omit the details. We obtain the surfaces [b] and [c] in diagram (b) of Figure 13.

It is straightforward to see (cf. [4]) that the dotted curve at $\epsilon = 0$ represents a curve of singularly perturbed homoclinic orbits with a generic unfolding. Note that this curve results from the back transformation of the curve $\hat{k}(\bar{\lambda})$ (cf. (c) in Figure 12) which represents points of this type for the blown-up system (5.9). The transformation (5.8) shows that the dotted curve of singularly perturbed homoclinic orbits of the system (3.7) is given by

$$k = \frac{c_6}{c_5}\lambda\cdot\hat{k}(\sqrt{\lambda/c_5}) =: k^{Ho}(\lambda) \qquad \text{with} \qquad k_\lambda^{Ho}(0) = \frac{c_6}{c_5}\cdot\bar{k}(4) = \frac{c_6}{c_5}\cdot\frac{\gamma}{7b_2}. \tag{5.39}$$

Comparing (5.39) and (5.37) we see that the inner border lines of the surfaces [a] and [d] are tangentially touching $k^{Ho}(\lambda)$ at the singularly perturbed Bogdanov point BP_s.

According to the theory we can only guarantee the surfaces [a], [b] and [c], [d] separated by the hatched gaps. But these gaps can be closed and there is one common surface according to numerical calculations. This surface intersects the (λ,k)-plane along the dotted curve $k^{Ho}(\lambda)$ of singularly perturbed homoclinic base points. In general we cannot expect the surface to be smooth with respect to ϵ along this curve. Note that during variation of $b_2 \to 0$ the curve tends to the k-axis and we approximate diagram (a).

5.5. Proof of Theorem 2.3. Using the derivatives in (5.39), (4.13) and the transformations (3.6), (5.8), (5.11) it is now possible to draw a complete picture of the singularly perturbed points of the pretransformed system (3.7) and the original system (2.5). Thus we can prove Theorem 2.3.

We obtain from (5.39) singularly perturbed homoclinic base points Ho_s at $[\lambda, k] = [\lambda, k^{Ho}(\lambda)]$ with $0 < \lambda \ll 1$ in the case of $c_5 = -f^0_{2x_1x_1} g^0_{x_1}/g^0_\lambda > 0$ and with $-1 \ll \lambda < 0$ in the case of $c_5 < 0$. The derivative at $\lambda = 0$ is given by $k^{Ho}_\lambda(0) = c_6\gamma/(c_5 7 b_2)$. The singularly perturbed saddle-nodes SN_s occur at $[\lambda, k] = [0, k]$, $0 < |k| \ll 1$ and finally we obtain singularly perturbed Hopf points HP_s at $k^H(\lambda)$ with $0 < \lambda \ll 1$ in the case of $\bar{H}'(0) = f^0_{2x_1x_1} g^0_\lambda/g^0_{x_1} > 0$ and with $-1 \ll \lambda < 0$ in the case of $\bar{H}'(0) < 0$. The corresponding derivative $k^H_\lambda(0)$ at $\lambda = 0$ is given by (4.13). Note that $\bar{H}'(0) \cdot c_5 = -(f^0_{2x_1x_1})^2 < 0$ so that the singularly perturbed Hopf points and the singularly perturbed homoclinic orbits exist on different sides of the k-axis which represents the singularly perturbed saddle-nodes.

The assumption of a generic unfolding of the singularly perturbed Bogdanov point ensures that the singularly perturbed points of lower degeneracy (saddle-nodes, Hopf points, homoclinic base points) are also generically unfolded with respect to (λ, ϵ) or (k, ϵ).

Qualitatively we obtain diagram (a) in Figure 2 with $(p_1, p_2) = (\lambda, k)$. Here we chose $c_5 > 0$, $k^{Ho}_\lambda(0) > 0$ and $k^H_\lambda(0) < 0$ which occurs in case of the model equation (1.5). The singularly perturbed points in the (p_1, p_2)-parameter space of the original system (2.5) are obtained by diffeomorphically deforming the curves according to the transformation (3.6).

In case of the Fitzhugh–Nagumo equation (1.3) we obtain diagram (b) in Figure 2; i.e., the singularly perturbed homoclinic orbits and the singularly perturbed Hopf points are both located on the a-axis. In particular they form a smooth curve at the singularly perturbed Bogdanov point at $a = 0$. This is a degenerate situation which results from the fact that the central u-system $\dot{u}_1 = u_2$, $\dot{u}_2 = -u_1^2 + u_1^3$ has a Bogdanov point in a Hamiltonian system.

In general we obtain this degenerate case if a certain normal form coefficient of the central system vanishes, as can be seen by the following argument. At $(\lambda, k, \epsilon) = 0$ the x-part of the system (3.7) assumes the following form after performing a standard normal form transformation:

$$\begin{aligned} \dot{x}_1 &= x_2 + O(|x|^3), \\ \dot{x}_2 &= \tfrac{1}{2} f^0_{2x_1x_1} \cdot x_1^2 + \mathrm{tr}_{x_1}(f_x)^0 \cdot x_1 x_2 + O(|x|^3). \end{aligned} \tag{5.40}$$

Hence the Bogdanov point is nondegenerate iff $f^0_{2x_1x_1} \neq 0$ and $\mathrm{tr}_{x_1}(f_x)^0 \neq 0$ (cf. [15]). The first inequality is valid according to (3.8). Concerning the second inequality the equations (5.39) and (4.13) yield the equivalence

$$k^{Ho}_\lambda(0) \neq k^H_\lambda(0) \qquad \Longleftrightarrow \qquad \mathrm{tr}_{x_1}(f_x)^0 \neq 0; \tag{5.41}$$

i.e., generically we have $\mathrm{tr}_{x_1}(f_x)^0 \neq 0$ and the singularly perturbed Bogdanov point causes curves of singularly perturbed homoclinic orbits, singularly perturbed saddle-nodes and singularly perturbed Hopf points which meet transversally at the origin. The corresponding (λ, k, ϵ)-unfolding diagram is depicted in diagram (a) of Figure 3. When approaching the degenerate case $\mathrm{tr}_{x_1}(f_x)^0 = 0$ of a nontransversal intersection the precise behavior of the surfaces in Figure 3 diagram (a) is not clear.

5.6. Numerical results. We close the paper by numerically continuing the surface of homoclinic orbits to the border line indicated by the open circles in Figure 3(a), Figure 9, and Figure 13(b). This is done for the model equation (1.5). With $(p_1, p_2) = (\lambda, k)$ and for fixed $-1 \ll k < 0$ the (λ, ϵ)-section through Figure 3(a) yields the structures depicted in Figure 14.

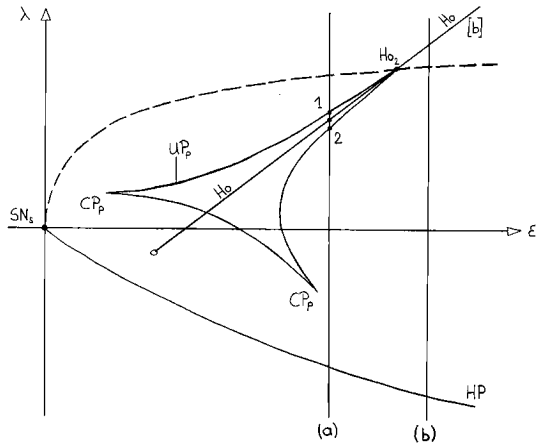
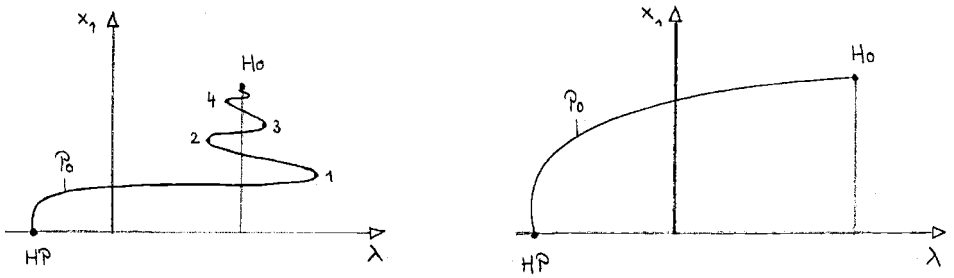


FIG. 14. *The* (λ, ϵ)*-section through Figure* 3(a) *for fixed* $-1 \ll k < 0$.



(a) (b)

FIG. 15. *The bifurcation diagrams along the* λ*-sections* (a) *and* (b) *in Figure* 14.

On the left side of the broken line we obtain three real eigenvalues at the stationary point $(u_1, u_2, v) = (\lambda, 0, -\lambda^2)$, whereas on the right side a conjugate-complex pair of eigenvalues $(\rho \pm i\omega) \in \mathbb{C}$ coexists with one real eigenvalue $\kappa \in \mathbb{R}$ with $|\kappa| > |\rho|$. Consequently, the homoclinic orbits Ho turn into homoclinic orbits of Shilnikov type [15] when crossing the broken line at the degenerate homoclinic orbit Ho_2 during variation of ϵ towards zero. Ho_2 is a well-known codimension-two homoclinic bifurcation (see [6] for numerical and theoretical results). Figure 15 shows the corresponding bifurcation diagrams along the λ-sections (a) and (b) in Figure 14.

In particular we obtain in the left diagram an infinite sequence 1,2,3,4, ... of periodic limit points with respect to λ. The first two limit points 1 and 2 are also marked along the line (a) in Figure 14. Now the numerical continuation of these points within the (λ, ϵ)-space of Figure 14 yields the closed curve UP_p of periodic limit points; i.e., the limit points 1 and 2 are connected by a curve that emanates from the degenerate homoclinic orbit Ho_2 and which contains additionally the two cusp points CP_p of periodic orbits. These cusp points arise by the coalescence of two limit points.

An analogous connection exists between the next two limit points 3 and 4 in Figure 15 and so forth. We conjecture that the (λ, ϵ)-unfolding diagram in Figure 14

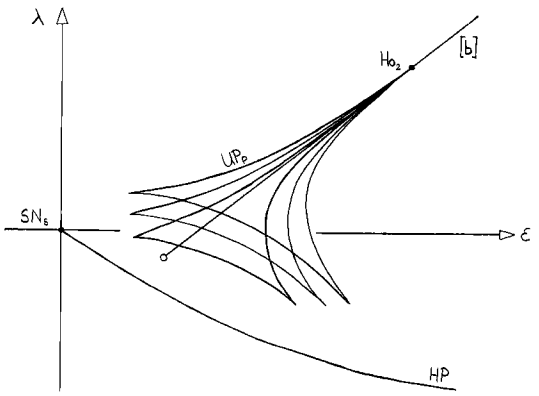


FIG. 16. *Closed curves of periodic limit points in the* (λ, ϵ)*-plane.*

can be extended by an infinite sequence of closed curves which approach the curve of homoclinic orbits Ho as indicated in Figure 16. Each of the closed curves represents a surface of periodic orbits that are connected in a rather complicated manner by limit points of periodic orbits.

Unfortunately, this observation does not clarify the behavior of the homoclinic orbits near the open circle. However, numerical calculations suggest that a canard like blow-up occurs as indicated in diagram (d) of Figure 3; i.e., the homoclinic orbits of Shilnikov type blow-up along the continuum of stationary points $v = -u_1^2$ which exists at $\epsilon = 0$. These blow-up phenomena typically occur in an exponentially small strip that is hardly accessible to numerical calculations.

Note that within the complete (λ, k, ϵ)-unfolding space of the model equation (1.5) we obtain a curve of open circles that passes through the singularly perturbed Bogdanov point at $k = 0$ (cf. Figure 3(a)). The (k, ϵ)-dependence seems to be of cubic order; i.e., we expect $k(\epsilon) = O(\epsilon^3)$ near the singularly perturbed Bogdanov point. This observation is purely motivated by numerical experiments. We did not succeed to predict it from formal asymptotic expansions.

Acknowledgment. This paper represents an extended part of the authors thesis. It was supported by the DFG–Schwerpunktprogramm "Ergodentheorie, Analysis und effiziente Simulation dynamischer Systeme" and was carried out during a four-month visit at the University of Bielefeld and a seven-month visit at the University of Tübingen. The author wishes to express his thanks to Prof. W.-J. Beyn, Prof. E. Bohl, and Prof. K. P. Hadeler.

REFERENCES

[1] V. I. Arnold, V. S. Afrajmovich, Y. S. Il'yashenko, and L. P. Shil'nikov, *Dynamical Systems* V: *Bifurcation Theory and Catastrophe Theory*, Springer-Verlag, Berlin, 1994.

[2] W.-J. Beyn, *The numerical computation of connecting orbits in dynamical systems*, IMA J. Numer. Anal., 9 (1990), pp. 379–405.

[3] W.-J. Beyn, *Numerical analysis of homoclinic orbits emanating from a Takens-Bogdanov point*, IMA J. Numer. Anal., 14 (1994), pp. 381–410.

[4] W.-J. Beyn and M. Stiefenhofer, *A direct approach to homoclinic orbits in the fast dynamics of singularly perturbed systems*, J. Dynam. Differential Equations, 11 (1997), pp. 671–709.

[5] R. G. Casten, H. Cohen, and P. A. Lagerstrom, *Perturbation analysis of an approximation to the Hodgkin–Huxley theory*, Quart. Appl. Math., 32 (1975), pp. 365–402.

[6] A. R. Champneys and Y. A. Kuznetsov, *Numerical detection and continuation of codimension-two homoclinic bifurcations*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 4 (1994), pp. 785–822.
[7] S.-N. Chow, J. Hale, and J. Mallet-Paret, *An example of bifurcation to homoclinic orbits*, J. Differential Equations, 37 (1980), pp. 351–373.
[8] M. Diener, *Étude générique des canards*, IRMA, Strasbourg, France, 1981.
[9] E. J. Doedel and J. P. Kernevez, *AUTO – Software for continuation and bifurcation problems in ordinary differential equations*, Appl. Math., Technical Report, CALTECH, 1986.
[10] F. Dumortier and R. Roussarie, *Canard Cycles and Center Manifolds*, Mem. Amer. Math. Soc. 121, AMS, Providence, RI, 1996.
[11] W. Eckhaus, *Relaxation Oscillations Including a Standard Chase on French Ducks*, Lecture Notes in Math. 985, Springer-Verlag, Berlin, New York, 1983.
[12] N. Fenichel, *Geometric singular perturbation theory for ordinary differential equations*, J. Differential Equations, 31 (1979), pp. 53–98.
[13] C. G. Gibson, *Singular Points of Smooth Mappings*, Pitman, Boston, 1979.
[14] J.-L. Martiel and A. Goldbeter, *A model based on receptor desensitization for cyclic AMP signalling in Dictyostelium cells*, J. Biophs., 52 (1987), pp. 802–828.
[15] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Appl. Math. Sci. 42, Springer-Verlag, New York, 1990.
[16] S. P. Hastings, *Single and multiple pulse waves for the FitzHugh–Nagumo equations*, SIAM J. Appl. Math., 42 (1982), pp. 247–260.
[17] H. Haken, *Synergetics. An Introduction*, Springer-Verlag, Berlin, New York, 1978.
[18] C. K. R. T. Jones and N. Kopell, *Tracking invariant manifolds with differential forms in singularly perturbed systems*, J. Differential Equations, 108 (1994), pp. 64–88.
[19] M. Koper, *Bifurcations of mixed-mode oscillations in a three-variable autonomous van der Pol–Duffing model with a cross-shaped phase diagram*, Phys. D., 80 (1995), pp. 72–94.
[20] Y. Kuznetsov, *Elements of Applied Bifurcation Theory*, Springer-Verlag, New York, 1995.
[21] J. D. Murray, *Mathematical Biology*, Springer-Verlag, Berlin, New York, 1989.
[22] NAG (Numerical Algorithms Group), Oxford, 1983.
[23] K. J. Palmer, *Exponential dichotomies and transversal homoclinic points*, J. Differential Equations, 55 (1984), pp. 225–256.
[24] M. Shub, A. Fathi, and R. Langevin, *Global Stability of Dynamical Systems*, Springer-Verlag, New York, Berlin, 1987.
[25] M. Stiefenhofer, *Singuläre Störung und Verzweigung bei Schleimpilzen*, Ph.D. thesis, University of Konstanz, Hartung-Gorre Verlag, Konstanz, Germany, 1995.
[26] M. Stiefenhofer, *Singular perturbation with limit points in the fast dynamics*, Z. Angew. Math. Phys., 49 (1998), pp. 730–758.
[27] M. Stiefenhofer, *Singular perturbation with Hopf points in the fast dynamics*, Z. Angew. Math. Phys., 49 (1998), pp. 602–629.
[28] P. Szmolyan, *Transversal heteroclinic and homoclinic orbits in singular perturbation problems*, J. Differential Equations, 92 (1991), pp. 252–281.
[29] A. B. Vasil'eva, *Asymptotic behaviour of solutions to certain problems involving nonlinear differential equations containing a small parameter multiplying the highest derivatives*, Russian Math. Surveys, 18 (1963), pp. 13–84.

SIAM J. MATH. ANAL.
Vol. 32, No. 4, pp. 854–869

OPTIMAL DESIGN AND CONSTRAINED QUASICONVEXITY*

PABLO PEDREGAL†

Abstract. We propose an alternative formulation for a typical optimal design problem in dimension two that allows for dependence on derivatives of the state. By means of a local formulation, we relate the optimal design problem to an equivalent vector variational problem. Relaxation in this framework involves the notion of constrained quasiconvexity.

Key words. optimal design, vector variational problems, quasiconvexity

AMS subject classifications. 49J20, 49J45, 74P10

PII. S0036141099356507

1. Introduction.

We would like to consider the following typical optimal design problem where the cost functional has the form

$$\overline{I}(g) = \int_\Omega W(x, g(x), w(x), \nabla w(x))\, dx,$$

the design variable g is constrained in some way and is related to w through the diffusion equation

$$-\operatorname{div}(g\nabla w) = f \quad \text{in} \quad H^{-1}(\Omega), \tag{1.1}$$

where $w \in H_0^1(\Omega)$, and $f \in H^{-1}(\Omega)$ is known. Ω is assumed to be a domain in $\mathbf{R}^2$ and the density $W : \Omega \times \mathbf{R} \times \mathbf{R} \times \mathbf{R}^2 \to \mathbf{R}^*$ is supposed to enjoy typical regularity properties, essentially continuity with respect to g, w, and ∇w and measurability with respect to x. $\mathbf{R}^*$ stands for $\mathbf{R} \cup \{+\infty\}$. No convexity whatsoever is assumed on W. Typical optimal design problems ask for constraints on g of the type $g(x) \in \{a, b\}$ for each $x \in \Omega$, where $0 < a < b$ and

$$\frac{1}{|\Omega|}\int_\Omega g(x)\, dx = \lambda a + (1-\lambda)b \tag{1.2}$$

for some fixed $\lambda \in (0, 1)$. We will focus in this work on this case as well, although other situations can also be treated. In precise terms, we will be concerned with the problem

$$\text{Minimize} \quad \overline{I}(g) = \int_\Omega W(x, g(x), w(x), \nabla w(x))\, dx,$$

subject to

$$g(x) \in \{a, b\} \text{ for almost everywhere (a.e.) } x \in \Omega, \quad \frac{1}{|\Omega|}\int_\Omega g(x)\, dx = \lambda a + (1-\lambda)b,$$
$$-\operatorname{div}(g\nabla w) = f, \quad w \in H_0^1(\Omega),$$

*Received by the editors May 21, 1999; accepted for publication (in revised form) July 24, 2000; published electronically December 13, 2000. This work has been supported by PB96-0534 of the DGES (Spain).

http://www.siam.org/journals/sima/32-4/35650.html

†ETSI Industriales, Universidad de Castilla-La Mancha, 13071 Ciudad Real, Spain (ppedrega @ind-cr.uclm.es).

where $0 < a < b$, $\lambda \in (0,1)$, $f \in H^{-1}(\Omega)$, and $W : \Omega \times \mathbf{R} \times \mathbf{R} \times \mathbf{R}^2 \to \mathbf{R}^*$ are given.

We would like in our analysis to link these optimal design problems to questions about vector variational problems, quasiconvexity, gradient Young measures, etc., just as in the pioneering work [9]. See also [1]. We will actually use some basic, important tools that played a role in these works. The main point in our paper is to examine and identify the underlying variational structure of this class of structural design problems, particularly what relates to relaxation. The novelty lies in the explicit dependence of the cost functional with respect to the gradient variable ∇w. Obviously, the analysis we are about to carry out for this situation can also be applied to the particular case of no dependence on this variable. For some particular choices of W, these problems have been considered in [17], although treated from a different perspective.

For the convenience of the reader, and in order to stress the main point we would like to convey, let us consider a simplified, one-dimensional situation [14]

$$\text{Minimize} \quad \overline{I}(g) = \int_0^1 W(x, g(x), w(x), w'(x))\, dx,$$

subject to

$$g(x) \in \{a, b\} \text{ for a.e. } x \in (0,1), \quad \int_0^1 g(x)\, dx = \lambda a + (1-\lambda) b,$$
$$-(g(x)w'(x))' = f, \quad w(0) = w(1) = 0,$$

where $0 < a < b$, $\lambda \in (0,1)$, $f \in H^{-1}(0,1)$ and $W : (0,1) \times \mathbf{R}^3 \to \mathbf{R}^*$ are given. Because of the restrictions on g we can write, for a suitable characteristic function χ,

$$g(x) = \chi(x) a + (1 - \chi(x))\, b, \quad \int_0^1 \chi(x)\, dx = \lambda.$$

Moreover, if

$$W_t(x, u, v) = W(x, t, u, v), \quad t = a, b,$$

the cost functional can be thought of as depending on χ:

$$I(\chi) = \int_0^1 \left[\chi(x) W_a(x, w(x), w'(x)) + (1 - \chi(x))\, W_b(x, w(x), w'(x))\right] dx.$$

On the other hand, if $-U'' = f$, $U \in H_0^1(0,1)$, then $-(gw' + U')' = 0$ and thus $g(x)w'(x) + U(x)$ must be constant throughout $(0,1)$. If we define a density $\varphi : (0,1) \times \mathbf{R}^3 \to \mathbf{R}^*$ by putting

$$\varphi(x, w, w', k) = \begin{cases} W_a(x, w, w'), & aw'(x) + U'(x) = k, \\ W_b(x, w, w'), & bw'(x) + U'(x) = k, \\ +\infty & \text{else}, \end{cases}$$

then our optimal design criterion can be written in terms of φ by minimizing the integral

$$J(w, k) = \int_0^1 \varphi(x, w(x), w'(x), k)\, dx,$$

where $w \in H_0^1(0,1)$ and k must be constant in $(0,1)$. The whole point of this formulation is that w and k in the pairs (w,k) are not related to each other in any way because the differential, nonlocal constraint has been recorded in the fact that k must be constant. The analysis would proceed by examining this equivalent variational problem. This is the type of analysis we would like to carry out for the two-dimensional situation.

It is by now a well-established fact that these optimal design problems lack classical solutions [12]. If we call χ, as in the previous one-dimensional example, the characteristic function of the set where $g = a$, then the cost functional $\overline{I}$ can also be interpreted in terms of χ. Optimizing sequences of such characteristic functions behave in such a way that there is no hope to obtain a new characteristic function as a limit in any sense, so that there will be no solution in the class of characteristic functions. This sort of behavior is typical of nonconvex variational problems. The main goal of the paper is to examine the relationship between these optimal design problems and nonconvex vector variational problems. In this regard, our point of view is similar to the one in [9], [10], [11] as pointed out in the previous paragraph.

The main feature of this type of optimization problems is the nonlocality expressed through the differential constraint (1.1). The difficulties associated with this issue can be in part overcome by the observation that in dimension two divergence-free fields and curl-free fields are essentially the same when the domain Ω is simply connected, which is an assumption we impose in our domain. Hence if $U \in H_0^1(\Omega)$ is the solution of the problem

$$-\Delta U = f \quad \text{in} \quad H^{-1}(\Omega),$$

then

$$\operatorname{div}(-g\nabla w + \nabla U) = 0.$$

Therefore if T is a $\pi/2$-rotation, there exists a stream function $v \in H^1(\Omega)$ such that

$$\nabla U = g\nabla w + T\nabla v, \tag{1.3}$$

provided Ω is simply connected [7]. This identity ties g, w, and v with U, a data of the problem. Thus we can think of the pair $u = (w, v)$, a vector field in $\mathbf{R}^2$, as our "independent" variable and of g, our design variable, as determined by (1.3). The idea is to rewrite our original cost functional $\overline{I}$ as a local functional in terms of u and carry out an analysis of the problem in this new formulation. Obviously, we have to pay close attention to the fact that (1.3) must force g to take on the values a and b exclusively, and at the same time we also have to enforce the volume constraint (1.2). This same approach was used in [15] for an analysis of magnetostriction.

In the present case, we show the equivalence of the above design problem and the following vector variational problem:

$$\inf\left\{\int_\Omega \varphi(x,u(x),\nabla u(x))\,dx : u \in H^1(\Omega;\mathbf{R}^2), u^{(1)} \in H_0^1(\Omega),\right.$$
$$\left.\int_\Omega \psi(x,u(x),\nabla u(x))\,dx = \lambda\right\}$$

for appropriate densities φ and ψ. See section 2 in order to understand the relationship between φ, ψ and W. $u^{(1)}$ indicates the first component of u. We analyze relaxation under this new formulation and show the existence of an energy density

$$\Psi(x,u,F,t) : \Omega \times \mathbf{R}^2 \times \mathbf{M}^{2\times 2} \times \mathbf{R} \to \mathbf{R}^*$$

which is continuous with respect to (u, F, t) whenever it is finite, and measurable with respect to x, that enjoys the joint convexity property with respect to the pairs (F, t) [5], [6]

$$\Psi(x, u, F, t) \leq \frac{1}{|\Omega|} \int_\Omega \Psi(x, u, F + \nabla V(y), t + \theta(y))\, dy$$

for any $V \in W_0^{1,\infty}(\Omega; \mathbf{R}^2)$, and $\theta \in L^\infty(\Omega)$ with $\int_\Omega \theta(y)\, dy = 0$. This is in fact the key constitutive assumption on Ψ to ensure an optimal solution of the problem

$$\inf \left\{ \int_\Omega \Psi(x, u(x), \nabla u(x), t(x))\, dx : u \in H^1(\Omega; \mathbf{R}^2), u^{(1)} \in H_0^1(\Omega), \right.$$
$$\left. 0 \leq t(x) \leq 1, \int_\Omega t(x)\, dx = \lambda\, |\Omega| \right\}.$$

This optimal solution encodes minimizing sequences for the optimal design problem. The relaxed energy density Ψ yields, for fixed (x, u), the optimal design corresponding to volume fraction t with respect to the value a, when the gradient of the state takes on the affine boundary values given by F. The vector nature of the problem reflects the nonlocality of the differential restriction (1.1).

In section 2, we precisely explain and show the equivalence of the initial optimal design problem and the associated vector variational principle as well as the relevant, relaxed integrand. Section 3 focuses on the analysis of this relaxed integrand, and in particular, we prove its joint convexity property which is the essential ingredient in order to have existence of solutions for the relaxed formulation. The last section deals with a typical relaxation result for the equivalent vector variational problem and incorporates a short discussion on the difficulties attached to the computation on explicit examples.

2. An alternative formulation. We describe how an equivalent, alternative formulation for our optimal design problem can be set up. This involves a vector variational principle to which all the typical ideas and techniques of the calculus of variations can be applied. The volume constraint, however, forces us to introduce the concept of constrained quasiconvexity and constrained quasiconvexification. The resulting envelope depends upon a gradient variable and a volume fraction.

In order to enforce the unique values a and b in the range of g from the outset, we write

$$g(x) = \chi(x)a + (1 - \chi(x))\, b,$$

χ being the characteristic function of a subset of Ω with

$$\int_\Omega \chi(x)\, dx = \lambda\, |\Omega|\,.$$

In this way, the cost functional can be thought of as depending on χ:

$$\bar{I}(\chi) = \int_\Omega \left[\chi(x) W_a(x, w(x), \nabla w(x)) + (1 - \chi(x)) W_b(x, w(x), \nabla w(x))\right]\, dx, \tag{2.1}$$

where W_a and W_b are appropriate, finite-everywhere energy densities and

$$\operatorname{div}(-\left[\chi a + (1 - \chi)b\right] \nabla w + \nabla U) = 0. \tag{2.2}$$

Indeed, $W_a(x,w,F) = W(x,a,w,F)$ and the analogous for b. Notice that χ is the characteristic function of the set where $\varphi = W_a$. U is taken as before: the solution of the problem

$$-\Delta U = f \quad \text{in } H^{-1}(\Omega), \quad U \in H_0^1(\Omega).$$

We have used the same letter $\overline{I}$ to designate the cost functional.

We have already pointed out in the introduction that

$$\nabla U = [\chi a + (1-\chi)b]\,\nabla w + T\nabla v \tag{2.3}$$

can be regarded as a substitute for the differential equation (2.2) if the domain Ω is simply connected. This equation is the clue for trying to recover the integral in (2.1) by means of a single cost density depending on the vector gradient $(\nabla w, \nabla v)$. In fact, we define a new energy density $\varphi : \Omega \times \mathbf{R}^2 \times \mathbf{M}^{2\times 2} \to \mathbf{R}^*$ by putting

$$\varphi(x,u,F) = \begin{cases} W_a(x,u^{(1)},F^{(1)}) & \text{if } aF^{(1)} + TF^{(2)} = \nabla U(x), \\ W_b(x,u^{(1)},F^{(1)}) & \text{if } bF^{(1)} + TF^{(2)} = \nabla U(x), \\ +\infty & \text{else.} \end{cases} \tag{2.4}$$

$F^{(i)}$, $i = 1,2$, designates the ith-row of F, and $u^{(i)}$ the ith component of u. For notational convenience, set for a real $t \in \mathbf{R}$ and a vector $B \in \mathbf{R}^2$

$$\Lambda(t,B) = \left\{ F \in \mathbf{M}^{2\times 2} : tF^{(1)} + TF^{(2)} = B \right\},$$

a two-dimensional linear manifold in the space of 2×2 matrices, $\mathbf{M}^{2\times 2}$, so that

$$\{\varphi < +\infty\} = \left\{ (x,u,F) \in \Omega \times \mathbf{R}^2 \times \mathbf{M}^{2\times 2} : F \in \Lambda(a,\nabla U(x)) \cup \Lambda(b,\nabla U(x)) \right\}.$$

There is, however, the ambiguity of how φ is defined on the intersection

$$\Lambda(a,\nabla U(x)) \cap \Lambda(b,\nabla U(x)) = \begin{pmatrix} 0 \\ T^{-1}\nabla U(x) \end{pmatrix}.$$

It is quite clear that in the set E where ∇w in (2.3) vanishes, we are free to choose a or b (i.e., $\chi = 1$ or $\chi = 0$, respectively) so as to minimize $W_a(x,w,0)$ or $W_b(x,w,0)$. Notice that w is constant a.e. in E. In this case, we would have to make the optimal choice

$$\min\left\{W_a(x,w,0), W_b(x,w,0)\right\}.$$

Hence, the definition of φ must incorporate this feature, and we redefine φ as

$$\varphi(x,u,F) = \begin{cases} W_a(x,u^{(1)},F^{(1)}) & \text{if } F \in \Lambda(a,\nabla U(x)) \setminus \Lambda(b,\nabla U(x)), \\ W_b(x,u^{(1)},F^{(1)}) & \text{if } F \in \Lambda(b,\nabla U(x)) \setminus \Lambda(a,\nabla U(x)), \\ \min\left\{W_a(x,u^{(1)},0), W_b(x,u^{(1)},0)\right\} & \text{if } F \in \Lambda(a,\nabla U(x)) \cap \Lambda(b,\nabla U(x)), \\ +\infty & \text{else.} \end{cases} \tag{2.5}$$

On the other hand, the volume constraint

$$\int_\Omega \chi(x)\,dx = \lambda\,|\Omega|$$

will also play an important role and needs to be taken into account in our new variational principle. If we let

$$\psi(x,u,F)=\begin{cases}1/|\Omega| & \text{if } F\in\Lambda(a,\nabla U(x)),\\ +\infty & \text{else,}\end{cases} \tag{2.6}$$

we must ask for the condition

$$\int_\Omega \psi(x,u(x),\nabla u(x))\,dx=\lambda.$$

Set

$$I(u)=\int_\Omega \varphi(x,u(x),\nabla u(x))\,dx,$$
$$J(u)=\int_\Omega \psi(x,u(x),\nabla u(x))\,dx.$$

We can summarize the previous discussion in the following proposition.

PROPOSITION 2.1. *Let $u\in H^1(\Omega;\mathbf{R}^2)$ such that $u^{(1)}\in H^1_0(\Omega)$ and $I(u)<+\infty$. There exists a subset of Ω whose characteristic function χ verifies*

$$\int_\Omega \chi(x)\,dx=\lambda\,|\Omega|\,,\quad \overline{I}(\chi)=I(u).$$

Conversely, if χ is the characteristic function of a subset of Ω such that

$$\int_\Omega \chi(x)\,dx=\lambda\,|\Omega|\,,$$

there exists $u\in H^1(\Omega;\mathbf{R}^2)$ such that $u^{(1)}\in H^1_0(\Omega)$ and

$$I(u)=\overline{I}(\chi),\quad J(u)=\lambda.$$

We pretend to examine the variational problem

$$\inf\left\{I(u):u\in H^1(\Omega),u^{(1)}\in H^1_0(\Omega),J(u)=\lambda\right\}. \tag{2.7}$$

It is quite apparent that in defining φ as in (2.5), we have destroyed all the nice properties (φ is no longer a Carathéodory function, not even upper semicontinuous) that allow a treatment in the context of the calculus of variations. ψ is not a Carathéodory function either. They are still weak lower semicontinuous with respect to the gradient variable. We could introduce a multiplier to take care of the integral constraint $J(u)=\lambda$. But instead we propose to consider and analyze the envelope

$$\begin{aligned}\Psi(x,u,F,t)=\frac{1}{|\Omega|}\inf\Big\{&\int_\Omega \varphi(x,u,F+\nabla V(y))\,dy : V\in W^{1,\infty}_0(\Omega),\\ &\int_\Omega \psi(x,u,F+\nabla V(y))\,dy=t\Big\}.\end{aligned} \tag{2.8}$$

This is some kind of constrained quasiconvexification. In fact, the definition of Ψ in terms of an infimum over gradients is too rigid when the integrand may take the value $+\infty$ abruptly. We are willing to allow gradient Young measures to take the

place of the gradients ∇V in the definition of Ψ [8]. The next result specifies the main properties enjoyed by Young measures associated with minimizing sequences for the optimal design problem.

LEMMA 2.2. *If the coercivity*

$$c\left(|\nabla w|^2 - 1\right) \le W_t(x, w, \nabla w), \quad c > 0 \tag{2.9}$$

holds for $t = a$ and $t = b$, then the gradients $\{\nabla u_j\}$ of every minimizing sequence $\{u_j\}$ for (2.7) generate a H^1-Young measure $\nu = \{\nu_x\}_{x\in\Omega}$ such that

$$\operatorname{supp}(\nu_x) \subset \Lambda(a, \nabla U(x)) \cup \Lambda(b, \nabla U(x)),$$
$$\int_\Omega \nu_x \left(\Lambda(a, \nabla U(x))\right)\, dx = \lambda\, |\Omega|\,.$$

Proof. If $\{u_j\}$ is minimizing for I, then for a.e. $x \in \Omega$ we should have

$$(x, u_j(x), \nabla u_j(x)) \in \{\varphi < +\infty\}\,. \tag{2.10}$$

This is equivalent to saying

$$\nabla u_j(x) \in \Lambda(a, \nabla U(x)) \cup \Lambda(b, \nabla U(x))$$

for a.e. $x \in \Omega$.

On the one hand, the coercivity assumed in (2.9) implies that the sequence $\{\nabla u_j^{(1)}\}$ is bounded in $L^2(\Omega)$. Since U is a single function in $H^1(\Omega)$, the previous fact on the support of ∇u_j says that $\{\nabla u_j\}$ is truly bounded in $L^2(\Omega)$. By the fundamental existence theorem for Young measures [2], [13], this sequence (or rather an appropriate subsequence) generates a Young measure $\nu = \{\nu_x\}_{x\in\Omega}$. By subtracting suitable constants to $u_j^{(2)}$ (note that this does not change the value of the functional I) and noticing that $u_j^{(1)} \in H_0^1(\Omega)$, we can assume without loss of generality that $\{u_j\}$ is bounded in $H^1(\Omega)$ so that ν is a H^1-Young measure.

On the other hand, if $G(x, A)$ is any nonnegative, Carathéodory function such that $\{G(x, \nabla u_j(x))\}$ is weakly convergent in $L^1(\Omega)$ and

$$G(x, \cdot)|_{\Lambda(a,\nabla U(x))\cup\Lambda(b,\nabla U(x))} \equiv 0,$$

then clearly

$$0 = \lim_{j\to\infty} \int_\Omega G(x, \nabla u_j(x))\, dx = \int_\Omega \int_{\mathbf{M}^{2\times 2}} G(x, A)\, d\nu_x(A)\, dx.$$

The arbitrariness of G indicates the conclusion on the support of ν.

The same kind of argument for a function G such that

$$G(x, \cdot)|_{\Lambda(a,\nabla U(x))} \equiv 1$$

permits us to conclude that

$$\int_\Omega \nu_x \left(\Lambda(a, \nabla U(x))\right)\, dx = \lambda\, |\Omega|\,. \qquad \square$$

We then redefine the envelope in (2.8) by putting

$$\Psi(x,u,F,t)=\inf\left\{\int_{\mathbf{M}^{2\times 2}}\varphi(x,u,F+A)\,d\nu(A):\nu \text{ is a homogeneous } H^1\text{-Young measure}, \int_{\mathbf{M}^{2\times 2}}A\,d\nu(A)=0, \int_{\mathbf{M}^{2\times 2}}\psi(x,u,F+A)\,d\nu(A)=t\right\}. \tag{2.11}$$

In so doing, we are accepting sequences of gradients that might take on values off the admissible set $\Lambda(a,\nabla U(x))\cup\Lambda(b,\nabla U(x))$ in small parts of Ω. Strictly speaking the cost of such designs would be infinite despite the fact that the cost in terms of its underlying Young measure would be finite. However, the new definition of Ψ in terms of Young measures is, if anything, a lower bound for Ψ in (2.8). We find it quite reasonable to allow these generalized designs if we can lower the cost when looking at its associated Young measure. In those small sets, g would not take either of the two values a or b, but rather it would have to be interpreted as a matrix-valued function. A more precise way of expressing these ideas is to say that the associated generalized problem in terms of Young measures

$$\inf\left\{\int_\Omega\int_{\mathbf{M}^{2\times 2}}\varphi(x,u(x),A)\,d\nu_x(A)\,dx:\nu=\{\nu_x\}_{x\in\Omega} \text{ is a } H^1\text{-Young measure}, \int_{\mathbf{M}^{2\times 2}}A\,d\nu_x(A)=\nabla u(x), u\in H^1(\Omega;\mathbf{R}^2), u^{(1)}\in H_0^1(\Omega), \int_\Omega\int_{\mathbf{M}^{2\times 2}}\psi(x,u(x),A)\,d\nu_x(A)\,dx=\lambda\,|\Omega|\right\}, \tag{2.12}$$

might not have as a solution the Young measure associated with a minimizing sequence for the original optimal design problem. Our claim is that minimizers for this optimization problem are, if anything, better if we are willing to allow "small errors" in the design in the sense explained above. More precisely we can state the following proposition.

PROPOSITION 2.3. *Let m and $\tilde{m}$ be the infima in* (2.7) *and* (2.12)*, respectively. Under the coercivity on W_t assumed in Lemma* 2.2*, $\tilde{m}\le m$.*

This proposition is a direct consequence, for instance, of Theorem 6.11 in [13]. Notice that even though ψ is not globally continuous with respect to F, it is indeed continuous restricted to the support of ν by Lemma 2.2. φ, however, is not continuous, not even when restricted to the support of ν. The integrals of φ in the above infimum should be understood in the following sense:

$$\begin{aligned}\int_{\mathbf{M}^{2\times 2}}\varphi(x,u,F+A)\,d\nu(A)=&\int_{\Lambda(a,\nabla U(x))\setminus\Lambda(b,\nabla U(x))}\varphi(x,u,F+A)\,d\nu(A)\\&+\int_{\Lambda(b,\nabla U(x))\setminus\Lambda(a,\nabla U(x))}\varphi(x,u,F+A)\,d\nu(A)\\&+\varphi\left(x,u,\begin{pmatrix}0\\T^{-1}\nabla U(x)\end{pmatrix}\right)\nu\left(\begin{pmatrix}0\\T^{-1}\nabla U(x)\end{pmatrix}\right).\end{aligned}$$

The idea of "small errors" brought to mind above admits a more formal, rigorous treatment in the context of [17], [16], and [19]. We will not insist on this issue, since our aim is to examine a vector variational problem with integrand Ψ. From now on, we stick to the functions φ, ψ, and Ψ as defined in (2.5), (2.6), and (2.11), respectively.

3. Relaxed integrand. The main property of the envelope Ψ is its jointly convex nature [5], [6] on the pairs (F,t) under which we can show existence of optimal solutions. This jointly convex property can be formulated in a standard way by requiring

$$\Psi(x,u,F,t) \le \frac{1}{|\Omega|}\int_\Omega \Psi(x,u,F+\nabla V(y), t+\theta(y))\,dy \tag{3.1}$$

whenever $V \in W_0^{1,\infty}(\Omega)$, and $\theta \in L^\infty(\Omega)$ with $\int_\Omega \theta(y)\,dy = 0$.

THEOREM 3.1. *The envelope Ψ is jointly convex in (F,t) for fixed (x,u) in the sense of* (3.1).

Proof. We divide the proof in several steps.

Step 1. V is piecewise affine and θ is piecewise constant. Let $V \in W_0^{1,\infty}(\Omega)$ be a piecewise affine function, so that we can write

$$\nabla V = \sum_i F_i \chi_{\Omega_i},$$

where $\{\Omega_i\}$ are pairwise disjoint open subdomains with

$$|\Omega - \cup_i \Omega_i| = 0.$$

Let

$$\theta(y) = \sum_i t_i \chi_{\Omega_i}, \quad \sum_i t_i\,|\Omega_i| = 0.$$

If $\nu^{(i)} = \left\{\nu_y^{(i)}\right\}_{y\in\Omega_i}$ is admissible in (2.11) for $(F+F_i, t+t_i)$ and we define $\nu = \{\nu_y\}_{y\in\Omega}$ by

$$\int_{\mathbf{M}^{2\times 2}} G(A)\,d\nu_y(A) = \int_{\mathbf{M}^{2\times 2}} G(F_i+A)\,d\nu_y^{(i)}(A),$$

whenever G is continuous and $y \in \Omega_i$, then ν is a H^1-Young measure (see the characterization of gradient Young measures in [8]). Because its underlying deformation is

$$\begin{aligned}\int_{\mathbf{M}^{2\times 2}} A\,d\nu_y(A) =& \sum_i \chi_{\Omega_i}(y)\int_{\mathbf{M}^{2\times 2}} (F_i+A)\,d\nu_y^{(i)}(A)\\ =& \sum_i \chi_{\Omega_i}(y) F_i\\ =& \nabla V(y),\end{aligned}$$

and V vanishes on the boundary (in particular it has affine boundary values), we may consider its homogenized version $\overline{\nu}$ [13]. Then

$$\begin{aligned}\int_{\mathbf{M}^{2\times 2}} \psi(x,u,F+A)\,d\overline{\nu}(A) =& \frac{1}{|\Omega|}\int_\Omega\int_{\mathbf{M}^{2\times 2}} \psi(x,u,F+A)\,d\nu_y(A)\,dy\\ =& \frac{1}{|\Omega|}\sum_i \int_{\Omega_i}\int_{\mathbf{M}^{2\times 2}} \psi(x,y,F+F_i+A)\,d\nu_y^{(i)}(A)\,dy\\ =& \frac{1}{|\Omega|}\sum_i (t_i+t)\,|\Omega_i|\\ =& t.\end{aligned}$$

Therefore $\overline{\nu}$ is admissible in the optimization problem that defines $\Psi(x,y,F,t)$ and

$$\begin{aligned}|\Omega|\,\Psi(x,u,F,t) &\le \int_\Omega \int_{\mathbf{M}^{2\times 2}} \varphi(x,u,F+A)\,d\nu_y(A)\,dy \\ &= \sum_i \int_{\Omega_i} \int_{\mathbf{M}^{2\times 2}} \varphi(x,u,F+F_i+A)\,d\nu_y^{(i)}(A)\,dy.\end{aligned}$$

Due to the arbitrariness of each admissible $\nu^{(i)}$ in (2.11) for $(F+F_i,t+t_i)$, and the independence among themselves, we can conclude

$$|\Omega|\,\Psi(x,u,F,t) \le \sum_i \Psi(x,u,F+F_i,t+t_i)\,|\Omega_i|\,,$$

which is the joint convexity for (V,θ).

Step 2. $\Psi(x,u,F,\cdot)$ is convex as a function of t, and $\Psi(x,u,\cdot,t)$ is rank-one convex as a function of F. The first fact is a direct consequence of Step 1 by taking $V\equiv 0$. For the rank-one convexity, notice that $\nu = s\nu_1 + (1-s)\nu_2$ is admissible in (2.11) for the pair (F,t) if ν_i is admissible for (F_i,t), $i=1,2$, and $\operatorname{rank}(F_1-F_2)\le 1$ (see again [8]). As a direct consequence, the dependence of Ψ with respect to (F,t) is continuous wherever it is finite.

Step 3. If $\mathcal{F}(A)$ (for fixed (x,u)) designates the multifunction

$$\mathcal{F}(A) = \{t\in[0,1] : \Psi(x,u,A,t) < +\infty\}\,,$$

then

$$\begin{aligned}\mathcal{F}(A) &= (0,1) \quad \text{if } A\notin \Lambda(a,\nabla U(x))\cup\Lambda(b,\nabla U(x)),\\ \mathcal{F}(A) &= (0,1] \quad \text{if } A\in \Lambda(a,\nabla U(x)),\\ \mathcal{F}(A) &= [0,1) \quad \text{if } A\in \Lambda(b,\nabla U(x))\setminus\Lambda(a,\nabla U(x)).\end{aligned}$$

In particular, $\mathcal{F}$, as a multifunction, is lower semicontinuous: if $t\in\mathcal{F}(A)$ and $A_j\to A$, we can find $t_j\in\mathcal{F}(A_j)$ such that $t_j\to t$.

The interesting part of the previous statement concerns the computation of $\mathcal{F}(A)$ when A does not belong to either of the sets $\Lambda(a,\nabla U(x))$, $\Lambda(b,\nabla U(x))$. We will show that for any such matrix, and for any $t\in(0,1)$ there exists a laminate μ_A supported on $\Lambda(a,\nabla U(x))\cup\Lambda(b,\nabla U(x))$ with barycenter A and such that $\mu_A(\Lambda(a,\nabla U(x))) = t$. Since the structure of laminates is preserved by translation, and the sets $\Lambda(s,B)$ are two-dimensional linear manifolds of matrices, we can show our claim, without loss of generality, replacing $\nabla U(x)$ by 0.

We will proceed in two steps.

LEMMA 3.2. *If we denote by $A^{(i)}$, $i=1,2$, the rows of A, and*

$$(3.2)\quad \Big[\Big(aA^{(1)}+TA^{(2)}\Big)\cdot\Big(bA^{(1)}+TA^{(2)}\Big)\Big]^2 - 4ab\Big(A^{(1)}\cdot A^{(2)}\Big)^2 > 0, \quad A^{(1)}\cdot A^{(2)} < 0,$$

there exists a vector $Y\in\mathbf{R}^2$, and scalars $t>0$, $s\in\mathbf{R}$, such that

$$(3.3)\qquad A + \begin{pmatrix} Y \\ sY\end{pmatrix} \in \Lambda(a,0), \quad A - t\begin{pmatrix} Y \\ sY\end{pmatrix} \in \Lambda(b,0).$$

Proof. The conditions in (3.3) can be explicitly written as

$$-aA^{(1)} - TA^{(2)} = (a\mathbf{1} + sT)\, Y,$$
$$bA^{(1)} + TA^{(2)} = t\,(b\mathbf{1} + sT)\, Y,$$

where $\mathbf{1}$ is the identity matrix. After eliminating the vector Y, we obtain

$$(a\mathbf{1} + sT)(bA^{(1)} + TA^{(2)}) + t(b\mathbf{1} + sT)(aA^{(1)} + TA^{(2)}) = 0.$$

This identity will be possible if and only if

$$\begin{vmatrix} (a\mathbf{1} + sT)(bA^{(1)} + TA^{(2)}) \\ (b\mathbf{1} + sT)(aA^{(1)} + TA^{(2)}) \end{vmatrix} = 0. \tag{3.4}$$

Moreover, the fact that we should have $t > 0$ implies

$$(a\mathbf{1} + sT)(bA^{(1)} + TA^{(2)}) \cdot (b\mathbf{1} + sT)(aA^{(1)} + TA^{(2)}) < 0. \tag{3.5}$$

If we keep in mind that

$$\det \begin{vmatrix} x \\ Ty \end{vmatrix} = x \cdot y,$$

then, after a few elementary manipulations, we get that (3.4) can be written explicitly

$$A^{(1)} \cdot A^{(2)} s^2 - \left(aA^{(1)} + TA^{(2)}\right) \cdot \left(bA^{(1)} + TA^{(2)}\right) s + abA^{(1)} \cdot A^{(2)} = 0.$$

This quadratic equation will have two real solutions if and only if (3.2) holds. We claim that one of those two solutions is such that the additional requirement (3.5) is fulfilled. Checking this explicitly is almost impossible by hand.

To this aim we introduce two scalar functions $f(s)$ and $g(s)$ determined by

$$\begin{vmatrix} (a\mathbf{1} + sT)(bA^{(1)} + TA^{(2)}) \\ (b\mathbf{1} + f(s)T)(aA^{(1)} + TA^{(2)}) \end{vmatrix} = 0,$$

and

$$g(s) = (a\mathbf{1} + sT)(bA^{(1)} + TA^{(2)}) \cdot (b\mathbf{1} + f(s)T)(aA^{(1)} + TA^{(2)}).$$

The following properties of f and g are easy to show.

1. Under (3.2), f is a quotient of two nondegenerate linear functions. Notice that the fixed points of f are the solutions of our quadratic equation above (3.4).

2. $f(s) \to f_\infty$ as $s \to \pm\infty$, where f_∞ is such that

$$\begin{vmatrix} T(bA^{(1)} + TA^{(2)}) \\ (b\mathbf{1} + f_\infty T)(aA^{(1)} + TA^{(2)}) \end{vmatrix} = 0.$$

3. g never vanishes: $g(s) = 0$ requires $A \in \Lambda(a, 0)$ or $A \in \Lambda(b, 0)$ which is impossible by (3.2) (keep in mind the definition of $f(s)$).

4. g takes on positive and negative values: notice that

$$\frac{g(s)}{s} \to T(bA^{(1)} + TA^{(2)}) \cdot (b\mathbf{1} + f_\infty T)(aA^{(1)} + TA^{(2)}),$$

as $s \to \pm\infty$.

5. Since g is continuous whenever f is continuous, g changes signs when f is not defined (note that f being a quotient of two linear functions has one vertical asymptote).

6. It is elementary to check that the two fixed points of f lie in the right branch of its graph. Therefore, we must make sure that g is negative on this branch. Because of property 4 above, it suffices to enforce

$$T(bA^{(1)} + TA^{(2)}) \cdot (b\mathbf{1} + f_\infty T)(aA^{(1)} + TA^{(2)}) < 0.$$

After a few careful computations this inequality reduces to

$$(b-a)^2 A^{(1)} \cdot A^{(2)} + \frac{\left(aA^{(1)} + TA^{(2)}\right) \cdot \left(bA^{(1)} + TA^{(2)}\right)^2}{A^{(1)} \cdot A^{(2)}} < 0.$$

This is equivalent to requiring

$$A^{(1)} \cdot A^{(2)} < 0.$$

This proves our claim. □

An interesting way of expressing the conclusion of this result is by saying that for each matrix A verifying (3.2) there exists a gradient Young measure (a laminate) of the simple form

$$\mu_A = t(A)\delta_{A_a} + (1 - t(A))\delta_{A_b} \tag{3.6}$$

such that $A_a \in \Lambda(a, 0)$, $A_b \in \Lambda(b, 0)$, $t(A) \in [0, 1]$. Let us denote by Γ the set of matrices A satisfying (3.2).

We go back to the proof of Step 3. Let A be any matrix not belonging to either of the sets $\Lambda(a, 0)$ or $\Lambda(b, 0)$. We claim that there exists a vector Y such that

$$A + \begin{pmatrix} Y \\ -Y \end{pmatrix} \in \Lambda(a, 0), \quad A - t \begin{pmatrix} Y \\ -Y \end{pmatrix} \in \Gamma$$

for arbitrarily large values of $t > 0$.

The first condition forces us to take

$$Y = -(a\mathbf{1} - T)^{-1} \left(aA^{(1)} + TA^{(2)}\right).$$

Notice that $a\mathbf{1} - T$ is nonsingular and Y is a nonvanishing vector. Then

$$\frac{1}{t}\left(A - t\begin{pmatrix} Y \\ -Y \end{pmatrix}\right) \to \begin{pmatrix} Y \\ -Y \end{pmatrix}$$

as $t \to +\infty$, and this limit belongs to Γ trivially. This proves our claim. Replacing a by b we also have the same conclusion.

This remark enables us to consider the laminates

$$\nu_a = \frac{t}{1+t}\delta_{A+Y\otimes -Y} + \frac{1}{1+t}\mu_{A-tY\otimes -Y},$$
$$\nu_b = \frac{t}{1+t}\delta_{A+X\otimes -X} + \frac{1}{1+t}\mu_{A-tX\otimes -X},$$

according to (3.1), for appropriate vectors Y and X where,

$$A + Y \otimes \ -Y = A + \begin{pmatrix} Y \\ -Y \end{pmatrix} \in \Lambda(a,0),$$
$$A + X \otimes \ -X = A + \begin{pmatrix} X \\ -X \end{pmatrix} \in \Lambda(b,0).$$

Notice that both ν_a and ν_b have barycenter A. Since these two laminates are admissible in (2.11) and

$$\nu_a\left(\Lambda(a,0)\right) \geq \frac{t}{1+t}, \quad \nu_b\left(\Lambda(b,0)\right) \geq \frac{t}{1+t},$$

for arbitrarily large t, the computation of $\mathcal{F}(A)$ is finished in this case, keeping in mind that because $\Psi(x,u,A,\cdot)$ is convex, $\mathcal{F}(A)$ must be a subinterval of $[0,1]$.

What remains to be proved in Step 3 can be derived by exploiting the rank-one convexity of $\Psi(x,u,\cdot,t)$ shown in Step 2. Indeed, if A is a matrix in the union $\Lambda(a,0) \cup \Lambda(b,0)$ and F is a rank-one matrix such that $A+F$ and $A-F$ do not lie in the union of those two subspaces (nearly any such F is valid), then for $t \in (0,1)$

$$\Psi(x,u,A,t) \leq \frac{1}{2}\Psi(x,u,A+F,t) + \frac{1}{2}\Psi(x,u,A-F,t) < +\infty.$$

Finally, it is interesting to notice that

$$0 \notin \mathcal{F}(A)$$

when $A \in \Lambda(a,\nabla U(x)) \cap \Lambda(b,\nabla U(x))$. If $0 \in \mathcal{F}(A)$ for such A, then that would imply the existence of a nontrivial gradient Young measure supported entirely in $\Lambda(b,\nabla U(x)) \setminus \Lambda(a,\nabla U(x))$. By Theorem 4.1 in [3], the linear manifold $\Lambda(b,\nabla U(x))$ would contain at least one rank-one direction, which is not trivially the case.

Step 4. Approximation and conclusion. Let $V \in W_0^{1,\infty}(\Omega)$, and $\theta \in L^\infty(\Omega)$ with null average such that

$$\int_\Omega \Psi(x,u,F+\nabla V(y), t+\theta(y))\,dy < \infty.$$

Hence $t+\theta(y) \in \mathcal{F}(F+\nabla V(y))$ for a.e. $y \in \Omega$. It is well known that there exists a uniformly bounded sequence of piecewise affine functions V_j such that $V_j \to V$ strongly in every $W^{1,p}(\Omega)$, $p < \infty$ [4]. Let

$$\nabla V_j = \sum_i F_i^{(j)} \chi_{\Omega_i^{(j)}}.$$

By Step 3, and the a.e. convergence $F+\nabla V_j(y) \to F+\nabla V(y)$, we can find functions θ_j such that $t+\theta_j(y) \in \mathcal{F}(F+\nabla V_j(y))$ and $t+\theta_j(y) \to t+\theta(y)$ for a.e. $y \in \Omega$. Moreover, because of the structure of ∇V_j we can always choose

$$\theta_j = \sum_i \theta_i^{(j)} \chi_{\Omega_i^{(j)}}, \quad \sum_i \theta_i^{(j)} \left|\Omega_i^{(j)}\right| = 0,$$

where $t+\theta_i^{(j)} \in \mathcal{F}(F+F_i^{(j)})$. By Step 1,

$$\Psi(x,u,F,t) \leq \lim_{j\to\infty} \int_\Omega \Psi(x,u,F+\nabla V_j(y), t+\theta_j(y))\,dy.$$

The choice of the pairs $(\nabla V_j, \theta_j)$, which are uniformly bounded, has been made within the region where Ψ is continuous, so that by pointwise convergence and dominated convergence we can conclude the joint convexity property. □

4. Relaxation. Based on the joint convexity property for the integrand Ψ, we can now prove the following existence theorem.

THEOREM 4.1. *Assume*

$$c\left(|\nabla w|^2 - 1\right) \le W_t(x, w, \nabla w) \le C\left(|\nabla w|^2 + 1\right), \quad 0 < c < C \tag{4.1}$$

for $t = a$ and $t = b$. Then

1. *the functional*

$$J(u,t) = \int_\Omega \Psi(x, u(x), \nabla u(x), t(x))\, dx$$

is weak lower semicontinuous in $H^1(\Omega; \mathbf{R}^2) \times L^\infty(\Omega)$;

2. *the variational problem*

$$\inf\left\{\int_\Omega \Psi(x, u(x), \nabla u(x), t(x))\, dx : u - u_0 \in H_0^1(\Omega; \mathbf{R}^2), \right.$$
$$\left. \|t\|_{L^\infty(\Omega)} \le M_0, \int_\Omega t(x)\, dx = t_0\, |\Omega| \right\}$$

admits a solution for any $u_0 \in H^1(\Omega; \mathbf{R}^2)$, $M_0 > 0$, and $t_0 \in [-M_0, M_0]$.

Proof. Notice that the bounds assumed on W_t imply the upper bound

$$\Psi(x, u, F, t) \le C\left(1 + |F|^2\right)$$

whenever Ψ is finite. The conclusion of the theorem is a direct consequence of Theorems 4.4 and 5.1 in [5]. The dependence of the integrand on u can be dealt with easily as in [13]. □

We finally show existence of solutions of our relaxed functional directly related to the original optimal design problem as stated at the end of the introduction. The only change with respect to Theorem 4.1 refers to the additional constraints coming from the restrictions on our optimal design problem.

THEOREM 4.2. *Under the same assumptions as in Theorem* 4.1*, the variational problem*

$$\inf\left\{\int_\Omega \Psi(x, u(x), \nabla u(x), t(x))\, dx : u \in H^1(\Omega; \mathbf{R}^2), u^{(1)} \in H_0^1(\Omega), \right.$$
$$\left. 0 \le t(x) \le 1, \int_\Omega t(x)\, dx = \lambda\, |\Omega| \right\}$$

admits an optimal solution for any $\lambda \in (0,1)$.

The proof is a simple application of the direct method. Observe that the constraints

$$u^{(1)} \in H_0^1(\Omega), \quad \int_\Omega t(x)\, dx = \lambda\, |\Omega|, \quad 0 \le t(x) \le 1$$

are preserved under weak convergence.

It is now standard to produce a nonhomogeneous Young measure minimizer by finding optimal solutions of the problem defining the envelope $\Psi(x, u(x), \nabla u(x), t(x))$ for a.e. $x \in \Omega$ [13].

The reader may feel a little bit disappointed to discover that the relaxed energy density Ψ is not computable analytically, not even for the simplest examples one can think of. At least such computation would require a deeper and a more quantitative analysis related to gradient Young measures supported in prescribed sets that must also meet the additional requirement of having a certain amount of its mass supported in a particular subset. As we know, this may be a hard question [3]. For instance, if

$$W(x,g,w,\nabla w)=|\nabla w|^2,$$

then

$$\varphi(x,F)=\begin{cases}\left|F^{(1)}\right|^2 & \text{if } F\in\Lambda(a,\nabla U(x))\cup\Lambda(b,\nabla U(x)),\\ +\infty & \text{else,}\end{cases}$$

and

$$\Psi(x,F,t)=\inf\left\{\int_{\mathbf{M}^{2\times 2}}|A|^2\,d\nu(A):\nu\in\mathcal{A}_{(x,F,t)}\right\},$$

where $\mathcal{A}_{(x,F,t)}$ consists of the set of homogeneous H^1-Young measures supported in $\Lambda(a,\nabla U(x))\cup\Lambda(b,\nabla U(x))$, with first moment F and such that $\nu(\Lambda(a,\nabla U(x)))=t$. The difficulty in computing Ψ is not in the form of φ but rather in the structure of the set of admissible Young measures $\mathcal{A}_{(x,F,t)}$. It would be interesting to analyze computationally this procedure, at least replacing gradient Young measures by laminates. We plan to do this in a subsequent paper. For this particular example important conclusions were derived in [18] by using different techniques.

The only example where further information can be drawn is the case of no dependence on derivatives, because in this case the qualitative analysis carried out in section 3 is enough to compute the density Ψ. Indeed, if we assume for simplicity that

$$W_a(x,w)\le W_b(x,w)\quad\text{for all } (x,w),\tag{4.2}$$

then it is elementary to obtain explicit expressions for φ and Ψ, namely,

$$\begin{gathered}\varphi\,(x,u,F)=W_a(x,u^{(1)}),\quad F\in\Lambda(a,\nabla U(x)),\\ \varphi\,(x,u,F)=W_b(x,u^{(1)}),\quad F\in\Lambda(b,\nabla U(x))\setminus\Lambda(a,\nabla U(x)),\\ \varphi=+\infty\quad\text{else.}\end{gathered}$$

Consequently, keeping in mind the conclusions of Step 3 in section 3, it is straightforward to arrive at

$$\begin{gathered}\Psi(x,u,F,t)=tW_a(x,u^{(1)})+(1-t)W_b(x,u^{(1)}),\quad t\in(0,1),\\ \Psi(x,u,F,t)=W_a(x,u^{(1)}),\quad F\in\Lambda(a,\nabla U(x)),t=1,\\ \Psi(x,u,F,t)=W_b(x,u^{(1)}),\quad F\in\Lambda(b,\nabla U(x))\setminus\Lambda(a,\nabla U(x)),t=0,\\ \Psi(x,u,F,t)=+\infty\quad\text{else.}\end{gathered}$$

Thus under (4.2) the optimal design problem would be "equivalent" to the variational problem

$$\text{Minimize}\quad\int_\Omega\left[t(x)W_a(x,u^{(1)}(x))+(1-t(x))W_b(x,u^{(1)}(x))\right]dx$$

subject to the constraints

$$u \in H^1(\Omega;\mathbf{R}^2), \quad u^{(1)} \in H^1_0(\Omega),$$
$$0 \le t(x) \le 1, \quad \int_\Omega t(x)\,dx = \lambda\,|\Omega|\,,$$
$$\{t=0\} \subset \{\nabla u \in \Lambda(b,\nabla U)) \setminus \Lambda(a,\nabla U))\}\,, \quad \{t=1\} \subset \{\nabla u \in \Lambda(a,\nabla U))\}\,.$$

Notice, however, that this variational problem may not have optimal solutions because coercivity assumptions on the gradient ∇u are not guaranteed.

Acknowledgments. I want to thank several referees for their criticism in improving this paper, and R. Lipton for some interesting suggestions.

REFERENCES

[1] G. Allaire, E. Bonnetier, G. Francfort, and F. Jouve, *Shape optimization by the homogenization method*, Numer. Math., 76 (1997), pp. 27–68.
[2] J. M. Ball, *A version of the fundamental theorem for Young measures*, in PDE's and Continuum Models of Phase Transitions, M. Rascle, D. Serre, and M. Slemrod, eds., Lecture Notes in Phys. 344, Springer-Verlag, Berlin, 1989, pp. 207–215.
[3] K. Battacharya, N. Firoozye, R. D. James, and R. V. Kohn, *Restrictions on microstructure*, Proc. Roy. Soc. Edinburgh Sect. A, 124 (1994), pp. 843–878.
[4] I. Ekeland and R. Temam, *Convex Analysis and Variational Problems*, North-Holland, Amsterdam, 1976.
[5] I. Fonseca, D. Kinderlehrer, and P. Pedregal, *Relaxation in $BV(\Omega,\mathbf{R}^p)\times L^\infty(w^*)$ of functionals depending on strain and composition*, in Boundary Value Problems for Partial Differential Equations and Applications, C. Baiocchi et al., eds., Masson et Cie, Paris, 1994, pp. 113–152.
[6] I. Fonseca, D. Kinderlehrer, and P. Pedregal, *Energy functionals depending on elastic strain and chemical composition*, Calc. Var. Partial Differential Equations, 2 (1994), pp. 283–313.
[7] V. Girault and P. A. Raviart, *Finite Element Methods for Navier-Stokes Equations*, Springer-Verlag, Berlin, 1986.
[8] D. Kinderlehrer and P. Pedregal, *Gradient Young measures generated by sequences in Sobolev spaces*, J. Geom. Anal., 4 (1994), pp. 59–90.
[9] R. V. Kohn and G. Strang, *Optimal design and relaxation of variational problems,* I, Comm. Pure Appl. Math., 39 (1986), pp. 113–137.
[10] R. V. Kohn and G. Strang, *Optimal design and relaxation of variational problems,* II, Comm. Pure Appl. Math., 39 (1986), pp. 139–182.
[11] R. V. Kohn and G. Strang, *Optimal design and relaxation of variational problems,* III, Comm. Pure Appl. Math., 39 (1986), pp. 353–377.
[12] F. Murat, *Contre-exemples pur divers problémes ou le contrôle intervient dans les coefficients*, Ann. Mat. Pura Appl. (4), 112 (1977), pp. 49–68.
[13] P. Pedregal, *Parametrized Measures and Variational Principles*, Birkhäuser, Basel, 1997.
[14] P. Pedregal, *Optimization, relaxation and Young measures*, Bull. Amer. Math. Soc. (N.S.), 36 (1999), pp. 27–58.
[15] P. Pedregal, *Relaxation in magnetostriction*, Calc. Var. Partial Differential Equations, 10 (2000), pp. 1–19.
[16] E. Polak and Y. Y. Wardi, *A study of minimizing sequences*, SIAM J. Control. Optim., 22 (1984), pp. 599–609.
[17] T. Roubícek, *Constrained optimization: A general tolerance approach*, Aplikace Matematiky, 35 (1990), pp. 99–128.
[18] L. Tartar, *Remarks on Optimal Design Problems*, CMU Preprint 94-NA-010, Carnegie Mellon University, Pittsburgh, PA, 1994.
[19] J. Warga, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.

SIAM J. MATH. ANAL.
Vol. 32, No. 4, pp. 870–886

POINTWISE ERROR ESTIMATES FOR RELAXATION APPROXIMATIONS TO CONSERVATION LAWS*

EITAN TADMOR† AND TAO TANG‡

Abstract. We obtain sharp *pointwise* error estimates for relaxation approximation to scalar conservation laws with piecewise smooth solutions. We first prove that the first-order partial derivatives for the perturbation solutions are uniformly upper bounded (the so-called Lip^+ stability). A one-sided interpolation inequality between classical L^1 error estimates and Lip^+ stability bounds enables us to convert a global L^1 result into a (nonoptimal) local estimate. Optimal error bounds on the weighted error then follow from the maximum principle for weakly coupled hyperbolic systems. The main difficulties in obtaining the Lip^+ stability and the optimal pointwise errors are how to construct appropriate "difference functions" so that the maximum principle can be applied.

Key words. conservation laws, error estimates, relaxation method, optimal convergence rate, one-sided interpolation inequality, maximum principle

AMS subject classifications. 35L65, 34C26, 65M10, 65M15

PII. S0036141098349492

1. Introduction.

Consider the following stiff relaxation system:

$$\begin{cases} u_t^\epsilon + v_x^\epsilon = 0, \\ \\ v_t^\epsilon + \alpha u_x^\epsilon = -\frac{1}{\epsilon}\Big(v^\epsilon - f(u^\epsilon)\Big), \qquad \epsilon > 0, \end{cases} \tag{1.1}$$

for $(x,t) \in \mathbf{R} \times (0,\infty)$. The initial conditions associated with the above system are

$$u^\epsilon(x,0) = u_0(x), \qquad v^\epsilon(x,0) = f(u_0(x)). \tag{1.2}$$

The system (1.1) can be regarded as a singular perturbation problem, and the solutions are expected to converge, as ϵ tends to zero, to the entropy solutions of the equilibrium equation

$$\begin{cases} u_t + f(u)_x = 0, \qquad v = f(u), \\ \\ u(x,0) = u_0(x), \qquad v(x,0) = f(u_0(x)). \end{cases} \tag{1.3}$$

The relaxation limit for 2×2 nonlinear systems of conservation laws was first studied by Liu [11], who justified some nonlinear stability criteria for diffusion waves, expansion waves, and traveling waves. A general mathematical framework was analyzed for the nonlinear systems by Chen, Levermore, and Liu [1]. Consult [12] for a bird's eye view of recent results in this direction.

The presence of relaxation mechanisms is widespread in both the continuum mechanics as well as the kinetic theory contexts. Relaxation is known to provide a subtle

*Received by the editors December 21, 1998; accepted for publication (in revised form) August 23, 2000; published electronically December 13, 2000. This research was supported in part by ONR grant N00014-91-J-1076, NSF grant DMS97-06827, and NSERC Canada grant OGP0105545. Part of this research was carried out while the second author was visiting UCLA.

http://www.siam.org/journals/sima/32-4/34949.html

†Department of Mathematics, UCLA, Los Angeles, CA 90095 (tadmor@math.ucla.edu).

‡Department of Mathematics, The Hong Kong Baptist University, Kowloon Tong, Hong Kong (tangtao@fisher.math.hkbu.edu.hk).

dissipative mechanism for discontinuities against the destabilizing effect of nonlinear response [11]. The relaxation models can be loosely interpreted as discrete velocity kinetic equations. The relaxation parameter, ϵ, plays the role of the mean free path and the system models the macroscopic conservation law. In that sense they are a discrete velocity analogue of the kinetic equations introduced by Perthame and Tadmor [15] and Lions, Perthame, and Tadmor [10].

The relaxation approximation can also be used to construct numerical approximations to the equilibrium conservation laws. In [6], Jin and Xin developed a class of first- and second-order nonoscillatory numerical schemes for the conservation law (1.3), based on the relaxation approximation (1.1). Since the relaxation approximation (1.1) is formally an $\mathcal{O}(\epsilon)$ perturbation to (1.3), they can compute (1.1) without resolving the computational grid to $\mathcal{O}(\epsilon)$. Indeed, in their final form, it is seen that the relaxation parameter in these relaxation schemes plays no role. In particular, their $\epsilon = 0$-limit in the first-order case coincides with the central Lax–Friedrichs scheme, and their $\epsilon = 0$-limit in the second-order version corresponds to the central scheme of Nessyahu and Tadmor [13]. The nonoscillatory central schemes introduced in [13] are based on staggered evolution of the reconstructed averages—a high-order sequel to the celebrated first-order Lax–Friedrichs (staggered) scheme. An extension of the high-resolution central scheme to multidimensional problems can be found in [5]. The central schemes are simple, efficient, stable, and enjoy the main advantage of avoiding costly (upwind) Riemann solvers. In this context, relaxation schemes offer yet another way to derive a whole class of high-resolution Riemann-solvers-free central schemes. The key is how to discretize the relaxation, as outlined in [6].

There have been many recent studies concerning the asymptotic convergence of the relaxation systems to the corresponding equilibrium conservation laws as the rate of the relaxation tends to zero. Most of these results deal with either large-time, nonlinear asymptotic stability or the zero relaxation limit for Cauchy problems. Tveito and Winther [18, 26] provided an $\mathcal{O}(\epsilon^{1/3})$-rate of convergence for some relaxation systems with nonlinear convection arising in chromatography. Katsoulakis and Tzavaras [7] introduced a class of relaxation systems, the contractive relaxation systems, and established an $\mathcal{O}(\sqrt{\epsilon})$ error bound in the case that the equilibrium equation is a scalar multidimensional one. The approaches in [7, 18, 26] are based on the extensions of Kruzhkov and Kuznetzov-type error estimates [9]. Kurganov and Tadmor [8] studied convergence and error estimates for a class of relaxation systems, including (1.1) and the one arising in chromatography, and concluded an $\mathcal{O}(\epsilon)$ order of convergence for scalar convex conservation laws. The novelty of their approach is the use of a weak Lip'-measure of the error, which allows them to obtain sharp error estimates.[1] For the relaxation system (1.1), Natalini [12] proved that the solutions to the relaxation system converges strongly to the unique entropy solution of (1.2) as $\epsilon \to 0$. Based on a general framework developed in [23, 25], the first-order rate of convergence for (1.1) is established in the case when its equilibrium solutions are piecewise smooth [24], which is an improvement on the $\mathcal{O}(\sqrt{\epsilon})$ error bounds [7, 8]. The boundary layer effect in the small relaxation limit to the equilibrium scalar conservation laws was investigated in [27]. The existence and uniqueness for the initial-boundary value problems are established.

The convergence and the rate of convergence mentioned above are mostly in the L^1 sense. It is understood that the L^1 error estimate is a global one, while in many practical cases we are interested in the *local* behavior of $u(x,t)$. Consequently, when

[1] Here and below, Lip' stands for the dual of Lip topology.

the error is measured by the L^1-norm, there is a loss of information due to the poor resolution of shock waves in $u(x,t)$. Several authors have investigated pointwise error estimates: For a system of conservation laws, Goodman and Xin [4] proved that the viscosity methods approximating piecewise smooth solutions with finitely many *noninteracting* shocks have a local $\mathcal{O}(\epsilon)$ error bound away from the shocks. A general convergence theory for one dimensional (1D) scalar convex conservation laws was developed by Tadmor and coauthors; see, e.g., [13, 19]. They proved that when measured in the weak Lip'-topology, the convergence rate of the viscous solution is of order $\mathcal{O}(\epsilon)$ in the case of rarefaction-free initial data and is of order $\mathcal{O}(\epsilon|\ln\epsilon|)$ in the general case. These weak Lip'-estimates are then converted into the usual L^1 error bounds of order one-half, and moreover, pointwise error estimates of order one-third, $\mathcal{O}(\epsilon^{1/3})$, are derived. Pointwise error analysis for finite difference methods to scalar and system of conservation laws is given recently by Engquist and Yu [3], Engquist and Sjogreen [2]. In [20], the authors provided the *optimal* pointwise convergence rate for the viscosity approximation. They used an innovative idea which enables them to convert a *global* L^1-error estimate into a *local* error estimate. Using this local error estimate and a bootstrap argument they proved that the viscosity approximation satisfies a *pointwise* error estimate of order $\mathcal{O}(\epsilon)$ for all but finitely many neighborhoods of shock discontinuities, each of width $\mathcal{O}(\epsilon)$. The previous results for the optimal order one convergence rates, in both L^1 and L^∞ spaces, are all based on a matching method and traveling wave solutions; see, e.g., [3, 4, 23]. The approach introduced in [20] does not follow the characteristics but instead makes use of the energy method, and hence can be extended to other types of approximate solutions, e.g., [21].

The question that we address in this paper is concerned with the rate of pointwise convergence for the relaxation approximation (1.1). The main purpose is to establish the *optimal* pointwise convergence. The proof of our results is based on two ingredients:

- a one-sided interpolation inequality between the L^1 error estimates and Lip^+ stability bounds; and
- a comparison theorem (the maximum principle) for weakly coupled hyperbolic systems.

In section 2, we review the preliminary results required for obtaining our error bounds. As mentioned earlier, the L^1 error bounds for the relaxation approximation have been established by several authors. A rigorous Lip^+ stability bounds for the relaxation approximation will be established in section 3. In section 4 we first consider the case when there is only one shock in the solutions of the equilibrium equation (1.3); i.e., the set of shock S consists of only one smooth curve. In this case, we show that

$$\operatorname{dist}(x,S)|u(x,t)-u^\epsilon(x,t)|\le C\epsilon. \tag{1.4}$$

It implies that $|u(x,t)-u^\epsilon(x,t)|\le C(h)\epsilon$ for (x,t) which are at least $\mathcal{O}(h)$ away from the set of shocks. The result (1.4) can be generalized to finitely many shocks with *possible collisions*. In the final section, we discuss the possible extensions of the results obtained in this work.

2. Preliminaries. Several useful results for the relaxation approximation will be reviewed in this section. We begin by introducing the subcharacteristic condition.

2.1. Subcharacteristic condition. The main stability criterion can be (formally) derived by using the Chapman–Enskog expansion for the stiff relaxation

system (1.1)

$$u_t^\epsilon + f(u^\epsilon)_x = \epsilon\Big((\alpha - f'(u^\epsilon)^2)u_x^\epsilon\Big)_x + \mathcal{O}(\epsilon^2). \tag{2.1}$$

The above equation will be of parabolic type under the following stability condition, i.e., the subcharacteristic condition [28]:

$$\alpha > f'(u^\epsilon)^2. \tag{2.2}$$

In a recent paper, Natalini [12] provided a rigorous analysis for (1.1) that leads to the subcharacteristic condition (2.2) under some assumptions on α and the initial data u_0. More precisely, we state his results as follows.

LEMMA 2.1. *If α in the relaxation equation* (1.1) *and the initial data u_0 in* (1.2) *satisfy*

$$\sqrt{\alpha} > M(N_0), \tag{2.3}$$

where N_0 and M_0 are defined by

$$\begin{cases} N_0 := \max\left(\|u_0\|_{L^\infty},\ \|f(u_0)\|_{L^\infty}\right), M(N_0) := \sup_{|\zeta|\le B(N_0)} |f'(\zeta)|, \textit{with} \\ \\ B(N_0) := 2N_0 + F(2N_0), F(N_0) := \sup_{|\zeta|\le N_0} |f(\zeta)|, \end{cases} \tag{2.4}$$

then the relaxation system (1.1) *with initial condition* (1.2) *satisfies the subcharacteristic inequality* (2.2). *Moreover, the solution (u^ϵ, v^ϵ) for* (1.1) *is uniformly bounded with respect to ϵ:*

$$|u^\epsilon(x,t)| \le B(N_0), |v^\epsilon(x,t)| \le \sqrt{\alpha}B(N_0), \textit{for} (x,t) \in \mathbf{R}\times(0,\infty). \tag{2.5}$$ ∎

Throughout this paper, we will assume that the condition (2.3) is satisfied. Under this assumption, the subcharacteristic inequality is guaranteed and will be used to establish the Lip^+ stability and the pointwise error bounds.

2.2. Global L^1 error bounds. The L^1-error analysis for the relaxation approximation method has been presented by several authors. For general data, an optimal L^1-rate can be found in [7, 8], for example. This optimal $\mathcal{O}(\sqrt{\epsilon})$ L^1-rate is overviewed in section 3.2, based on the Lip' approach taken in [8] (for a more general class of relaxation models). For piecewise-smooth data, the *optimal* $\sim \mathcal{O}(\epsilon)$ L^1-convergence rate was recently obtained by Teng [24]. We state his results as follows.

LEMMA 2.2. *Assume α in the relaxation equation* (1.1) *and the initial data u_0 in* (1.2) *satisfy the conditions stated in Lemma* 2.1. *Assume that the solutions to the scalar convex conservation law* (1.3) *are piecewise smooth. Let (u^ϵ, v^ϵ) be the solutions of the relaxation problems* (1.1)–(1.2). *Then the following error estimate holds:*

$$\sup_{0\le t\le T}\Big(\|u^\epsilon(\cdot,t) - u(\cdot,t)\|_{L^1(\mathbf{R})} + \|v^\epsilon(\cdot,t) - v(\cdot,t)\|_{L^1(\mathbf{R})}\Big) \le C(T)\epsilon\,|\ln\epsilon|, \tag{2.6}$$

where $v = f(u)$. If there is no initial central rarefaction wave and no new generated shocks, then the error bound is improved to

$$\sup_{0\le t\le T}\Big(\|u^\epsilon(\cdot,t) - u(\cdot,t)\|_{L^1(\mathbf{R})} + \|v^\epsilon(\cdot,t) - v(\cdot,t)\|_{L^1(\mathbf{R})}\Big) \le C(T)\epsilon. \tag{2.7}$$ ∎

We shall utilize these L^1 global error bounds to derive the pointwise error estimate (1.4). The order of the global L^1 error bounds will not affect the general $\mathcal{O}(\epsilon)$-pointwise result (1.4), but it will affect the choice of the distance function—see [20] for details. Thus improved L^1-error bounds lead to sharper description of the shock layer, with an optimal shock layer of size $\sim \mathcal{O}(\epsilon)$ corresponding to the piecewise-smooth cases (2.6) and (2.7).

2.3. An interpolation inequality. We let $\|\bullet\|_{Lip^+}$ denote the Lip^+-seminorm

$$\|w\|_{Lip^+} := \operatorname{ess\,sup}_{x\neq y}\left[\frac{w(x)-w(y)}{x-y}\right]^+,$$

where $[w]^+ = H(w)w$, with $H(\bullet)$ the Heaviside function. The following lemma is due to Nessyahu and Tadmor [14, section 2]; its proof can be found in [20].

LEMMA 2.3. *Assume that $z \in L^1 \cap Lip^+(\mathcal{I})$, and $w \in C^1_{loc}(x-\delta, x+\delta)$ for an interior x such that $(x-\delta, x+\delta) \subset \mathcal{I}$. Then the following estimate holds:*

(2.8)
$$|z(x)-w(x)| \le \text{Const}\cdot\left[\frac{1}{\delta}\|z-w\|_{L^1} + \delta\{\|z\|_{Lip^+(x-\delta,x+\delta)} + |w|_{C^1_{loc}(x-\delta,x+\delta)}\}\right].$$

In particular, if the size of the smoothness neighborhood for w can be chosen so that

$$\delta \sim \|z-w\|^{1/2}_{L^1(\mathcal{I})}\cdot(\|z\|_{Lip^+} + |w|_{C^1_{loc}})^{-1/2} \le \frac{1}{2}|\mathcal{I}|, \tag{2.9}$$

then the following estimate holds:

$$|z(x)-w(x)| \le \text{Const}\cdot\|z-w\|^{1/2}_{L^1(\mathcal{I})}\cdot\left[\|z\|_{Lip^+} + |w|_{C^1_{loc}(x-\delta,x+\delta)}\right]^{1/2}. \tag{2.10}$$

□

Thus (2.10) tells us that if the global L^1-error $\|z-w\|_{L^1}$ is small, then the pointwise error $|z(x)-w(x)|$ is also small wherever w_x is bounded. This does not require the C^1-boundedness of z—the weaker one-sided Lip^+ bound of z will suffice.

2.4. A comparison lemma. The following maximum principle for weakly coupled hyperbolic systems plays an important role in this work. Consider the following system:

$$\begin{cases} \partial_t u_1 + \lambda_1(x,t)\partial_x u_1 = \alpha_{11}(x,t)u_1 + \alpha_{12}(x,t)u_2 + \beta_1(x,t), \\ \\ \partial_t u_2 + \lambda_2(x,t)\partial_x u_2 = \alpha_{21}(x,t)u_1 + \alpha_{22}(x,t)u_2 + \beta_2(x,t), \end{cases} \tag{2.11}$$

with C^1 local speeds, $\lambda_i(\cdot)$, and low-order terms on the right involving bounded coefficients, $\alpha_{ij}(\cdot), \beta_i(\cdot)$, $1 \le i,j \le 2$. The following lemma (see, e.g., [16, Theorem 13]), provides sufficient conditions which guarantee that if the initial and boundary data prescribed for (2.11) is nonpositive, the solution remains nonpositive.

LEMMA 2.4. *Consider the Cauchy problem for the weakly coupled hyperbolic systems* (2.11) *in a domain $E := \mathbf{D}\times(0,T)$, subject to nonpositive initial and boundary conditions*

$$\begin{aligned} &u_1(x,0) \le 0, \quad u_2(x,0) \le 0, && x \in \mathbf{D}, \\ &u_1(x,t) \le 0, \quad u_2(x,t) \le 0, && (x,t) \in \partial\mathbf{D}\times(0,T). \end{aligned} \tag{2.12}$$

Assume that the coefficient functions in (2.11) *satisfy*

$$\begin{aligned} &\alpha_{12}(x,t) \geq 0, \quad \alpha_{21}(x,t) \geq 0, (x,t) \in E, \\ &\beta_1(x,t) \leq 0, \quad \beta_2(x,t) \leq 0, (x,t) \in E. \end{aligned} \tag{2.13}$$

Then the solution of (2.11) *remains nonpositive in later time:*

$$u_1(x,t) \leq 0, \qquad u_2(x,t) \leq 0 \qquad \textit{for} \quad (x,t) \in E. \qquad \square \tag{2.14}$$

For the proof, we note that thanks to (2.13), the nonpositive maximal values, $U_i(t) := \sup_x u_i(x,t)$ are majorized by the ODEs, $\dot{U}_i = \alpha_{ii} U_i(t) + \alpha_{ij} u_j(t) + \beta_i(t) \leq \alpha_{ii} U_i(t)$, and hence these maximal values cannot increase in time.

The two important results, the Lip^+ stability and the optimal pointwise error bounds are all based on the above lemma. The main difficulty is how to construct appropriate object functions u_1, u_2 so that above lemma can be suitably applied.

3. Lip^+ stability and local error bounds. In this section, we assume that f is strictly convex, i.e.,

$$f''(u) \geq \beta > 0 \qquad \text{for} \quad u \in \mathbf{R}, \tag{3.1}$$

and that u_0 is Lip^+-bounded,

$$\|u_0\|_{Lip^+} < \infty. \tag{3.2}$$

DEFINITION 3.1. *We say that* $\{u^\epsilon(x,t)\}_{\epsilon>0}$ *are* Lip^+*-stable if the following estimate is fulfilled:*

$$\|u^\epsilon(\cdot,t)\|_{Lip^+} \leq \|u_0\|_{Lip^+}, \qquad t \geq 0. \tag{3.3}$$

3.1. Lip^+ stability. We will show that the family $\{u^\epsilon(x,t)\}_{\epsilon>0}$ is Lip^+-stable. Assume first that $u_0 \in C_0^1(\mathbf{R})$. This implies, by the standard regularity theory for the semilinear hyperbolic problems, that $(u^\epsilon, v^\epsilon) \in C^1(\mathbf{R} \times (0,T))$ for some $T > 0$. Differentiating the equations (1.1) with respect to x gives

$$(u^\epsilon_x)_t + (v^\epsilon_x)_x = 0, \tag{3.4}$$

$$(v^\epsilon_x)_t + \alpha (u^\epsilon_x)_x = -\frac{1}{\epsilon}(v^\epsilon_x - f'(u^\epsilon) u^\epsilon_x). \tag{3.5}$$

By doing $\sqrt{\alpha} \times (3.4) + (3.5)$ and $\sqrt{\alpha} \times (3.4) - (3.5)$, the above system can be put in the following diagonal form:

$$\begin{aligned} &(\sqrt{\alpha} u^\epsilon_x + v^\epsilon_x)_t + \sqrt{\alpha}(\sqrt{\alpha} u^\epsilon_x + v^\epsilon_x)_x = -\frac{1}{\epsilon}(v^\epsilon_x - f'(u^\epsilon) u^\epsilon_x), \\ &(\sqrt{\alpha} u^\epsilon_x - v^\epsilon_x)_t - \sqrt{\alpha}(\sqrt{\alpha} u^\epsilon_x - v^\epsilon_x)_x = \frac{1}{\epsilon}(v^\epsilon_x - f'(u^\epsilon) u^\epsilon_x). \end{aligned}$$

Letting

$$\bar{p} = \sqrt{\alpha} u^\epsilon_x + v^\epsilon_x, \qquad \bar{q} = \sqrt{\alpha} u^\epsilon_x - v^\epsilon_x$$

and by using the above results yield

$$\begin{cases} \bar{p}_t + \sqrt{\alpha}\bar{p}_x = \frac{1}{2\epsilon}\left(\frac{f'(u^\epsilon)}{\sqrt{\alpha}} - 1\right)\bar{p} + \frac{1}{2\epsilon}\left(\frac{f'(u^\epsilon)}{\sqrt{\alpha}} + 1\right)\bar{q}, \\ \bar{q}_t - \sqrt{\alpha}\bar{q}_x = -\frac{1}{2\epsilon}\left(\frac{f'(u^\epsilon)}{\sqrt{\alpha}} - 1\right)\bar{p} - \frac{1}{2\epsilon}\left(\frac{f'(u^\epsilon)}{\sqrt{\alpha}} + 1\right)\bar{q}. \end{cases} \tag{3.6}$$

We further introduce the transformations

$$\begin{cases} p = \bar{p} - \Big(\sqrt{\alpha} + f'(u^\epsilon)\Big)\|u_0\|_{Lip^+}, \\ \\ q = \bar{q} - \Big(\sqrt{\alpha} - f'(u^\epsilon)\Big)\|u_0\|_{Lip^+}. \end{cases} \tag{3.7}$$

Applying the above transformations to (3.6) gives

$$\begin{aligned} p_t + \sqrt{\alpha}p_x = {} & \frac{1}{2\epsilon}\left(\frac{f'(u^\epsilon)}{\sqrt{\alpha}} - 1\right)p + \frac{1}{2\epsilon}\left(\frac{f'(u^\epsilon)}{\sqrt{\alpha}} + 1\right)q \\ & -f''(u^\epsilon)\Big(u_t^\epsilon + \sqrt{\alpha}u_x^\epsilon\Big)\|u_0\|_{Lip^+}, \end{aligned} \tag{3.8}$$

$$\begin{aligned} q_t - \sqrt{\alpha}q_x = {} & -\frac{1}{2\epsilon}\left(\frac{f'(u^\epsilon)}{\sqrt{\alpha}} - 1\right)p - \frac{1}{2\epsilon}\left(\frac{f'(u^\epsilon)}{\sqrt{\alpha}} + 1\right)q \\ & +f''(u^\epsilon)\Big(u_t^\epsilon - \sqrt{\alpha}u_x^\epsilon\Big)\|u_0\|_{Lip^+}. \end{aligned} \tag{3.9}$$

It follows from (1.1), $u_t^\epsilon + v_x^\epsilon = 0$, that

$$\begin{aligned} u_t^\epsilon + \sqrt{\alpha}u_x^\epsilon &= -v_x^\epsilon + \sqrt{\alpha}u_x^\epsilon = q + \Big(\sqrt{\alpha} - f'(u^\epsilon)\Big)\|u_0\|_{Lip^+}, \\ u_t^\epsilon - \sqrt{\alpha}u_x^\epsilon &= -v_x^\epsilon - \sqrt{\alpha}u_x^\epsilon = -p - \Big(\sqrt{\alpha} + f'(u^\epsilon)\Big)\|u_0\|_{Lip^+}. \end{aligned}$$

The above observation, together with (3.8) and (3.9), lead to

$$\begin{aligned} p_t + \sqrt{\alpha}p_x = {} & \frac{1}{2\epsilon}\left(\frac{f'(u^\epsilon)}{\sqrt{\alpha}} - 1\right)p + \frac{1}{2\epsilon}\left(\frac{f'(u^\epsilon)}{\sqrt{\alpha}} + 1 - 2\epsilon f''(u^\epsilon)\|u_0\|_{Lip^+}\right)q \\ & -f''(u^\epsilon)\Big(\sqrt{\alpha} - f'(u^\epsilon)\Big)\|u_0\|_{Lip^+}^2, \end{aligned} \tag{3.10}$$

$$\begin{aligned} q_t - \sqrt{\alpha}q_x = {} & -\frac{1}{2\epsilon}\left(\frac{f'(u^\epsilon)}{\sqrt{\alpha}} - 1 + 2\epsilon f''(u^\epsilon)\|u_0\|_{Lip^+}\right)p - \frac{1}{2\epsilon}\left(\frac{f'(u^\epsilon)}{\sqrt{\alpha}} + 1\right)q \\ & -f''(u^\epsilon)\Big(\sqrt{\alpha} + f'(u^\epsilon)\Big)\|u_0\|_{Lip^+}^2. \end{aligned} \tag{3.11}$$

It follows from the subcharacteristic condition (2.2) that (3.10)–(3.11) is a weakly coupled hyperbolic system and its coefficients satisfy the requirements in (2.12) provided that ϵ is sufficiently small. We now check the initial conditions. First checking $p(x,0)$,

$$\begin{aligned} p(x,0) &= \sqrt{\alpha}u_0' + f'(u_0)u_0' - (\sqrt{\alpha} + f'(u_0))\|u_0\|_{Lip^+} \\ &= (\sqrt{\alpha} + f'(u_0))(u_0' - \|u_0\|_{Lip^+}) \le 0. \end{aligned}$$

Similarly, we have

$$q(x,0) = (\sqrt{\alpha} - f'(u_0))(u_0' - \|u_0\|_{Lip^+}) \le 0.$$

Using Lemma 2.4, we obtain

$$p(x,t) \le 0, \qquad q(x,t) \le 0 \qquad \text{for} \quad (x,t) \in \mathbf{R} \times (0,T). \tag{3.12}$$

It follows from (3.7) that

$$u_x^\epsilon = \frac{1}{2\sqrt{\alpha}}(p+q) + \|u_0\|_{Lip^+}.$$

This identity, together with (3.12), yields

$$u_x^\epsilon \le \|u_0\|_{Lip^+},$$

which is the Lip^+ stability (3.3) for u^ϵ when it is smooth. Finally, we extend our result to general initial data by the following standard procedure:

$$u_0^\delta(x) := \int \psi_\delta(x-y)u_0(y)dy,$$

where ψ_δ is a compactly supported nonnegative unit mass mollifier,

$$\psi_\delta(x) = \frac{1}{\delta}\psi\left(\frac{x}{\delta}\right), \qquad \int_{-\infty}^{\infty} \psi_\delta(x)dx = 1.$$

It is obvious that if $\|u_0\|_{Lip^+} < \infty$, then $\|u_0^\delta\|_{Lip^+}$ is also bounded. Consider the 2×2 stiff relaxation system (1.1) with the smooth initial data

$$u^\epsilon(x,0) = u_0^\delta(x), \qquad v^\epsilon(x,0) = f(u_0^\delta(x)). \tag{3.13}$$

Using the above proof we know that there exists a $T > 0$ such that

$$\|u^{\epsilon,\delta}(\bullet,t)\|_{Lip^+} \le \|u_0^\delta\|_{Lip^+} \quad \text{for} \quad t \in (0,T), \tag{3.14}$$

where $u^{\epsilon,\delta}$ is one component of the solution to (1.1) and (3.13). Letting $\delta \to 0+$ in (3.14) gives

$$\|u^\epsilon(\bullet,t)\|_{Lip^+} \le \|u_0\|_{Lip^+} \qquad \text{for} \quad t \in (0,T).$$

By standard continuation arguments for time, we can extend the desired Lip^+ stability result (3.3) for u^ϵ to any finite time interval.

We summarize what we have shown by stating the following.

THEOREM 3.1. *Assume α in the relaxation equation* (1.1) *and the initial data u_0 in* (1.2) *satisfy the conditions stated in Lemma* 2.1. *Assume $f'' > 0$. Then the family of solutions $\{u^\epsilon(x,t)\}_{\epsilon>0}$, given by the relaxation system* (1.1) *and initial data* (1.2), *are Lip^+-stable. Moreover, the functions $\{\sqrt{\alpha}u^\epsilon + v^\epsilon\}_{\epsilon>0}$ and $\{\sqrt{\alpha}u^\epsilon - v^\epsilon\}_{\epsilon>0}$ are also Lip^+-stable.* □

3.2. Error estimates based on *Lip′* theory. Equipped with the Lip^+-stability, one can derive $O(\sqrt{\epsilon})$ L^1- and local error bounds using the Lip' theory presented in [19]. The case for a general family of relaxation models was outlined in [8]; here is a brief overview for the particular case of the relaxation model (1.1).

To begin with, we derive the modified equation satisfied by u^ϵ. Consider the second equation in (1.1),

$$v_t^\epsilon + \alpha u_x^\epsilon = -\frac{1}{\epsilon}(v^\epsilon - f(u^\epsilon)). \tag{3.15}$$

We differentiate with respect to x and use the first equation of (1.1), $v_x^\epsilon = -u_t^\epsilon$, to find

$$u_t^\epsilon + f(u^\epsilon)_x = \epsilon(v_t^\epsilon + \alpha u_x^\epsilon)_x. \tag{3.16}$$

The term on the right is the *truncation error*. The main result in [19, 14] shows that when measured in the Lip'-norm, the global error, $u^\epsilon - u$, is governed by the truncation+initial errors

$$\|u^\epsilon - u\|_{Lip'} \le \text{Const}\,[\epsilon\|(v_t^\epsilon + \alpha u_x^\epsilon)_x\|_{Lip'} + \|u_0^\epsilon - u_0\|_{Lip'}]\,. \tag{3.17}$$

In our case of (1.2), there is no initial error. To measure the Lip'-size of the truncation error, we proceed along the lines of [8, Example 3]: we differentiate (1.1) with respect to t, obtaining

$$(u_t^\epsilon)_t + (v_t^\epsilon)_x = 0, \tag{3.18}$$

$$(v_t^\epsilon)_t + \alpha(u_t^\epsilon)_x = -\frac{1}{\epsilon}\Big(v_t^\epsilon - f'(u^\epsilon)u_t^\epsilon\Big). \tag{3.19}$$

Performing $\sqrt{\alpha}\times$ (3.18) + (3.19) and $\sqrt{\alpha}\times$ (3.18) − (3.19), then the above system can be put in the following diagonal form in terms of the characteristic variables, $\bar r := \sqrt{\alpha}u_t^\epsilon + v_t^\epsilon$ and $\bar s := \sqrt{\alpha}u_t^\epsilon - v_t^\epsilon$,

$$\begin{cases} \bar r_t + \sqrt{\alpha}\bar r_x = \frac{1}{2\epsilon}\left(\frac{f'(u^\epsilon)}{\sqrt{\alpha}} - 1\right)\bar r + \frac{1}{2\epsilon}\left(\frac{f'(u^\epsilon)}{\sqrt{\alpha}} + 1\right)\bar s, \\ \bar s_t - \sqrt{\alpha}\bar s_x = -\frac{1}{2\epsilon}\left(\frac{f'(u^\epsilon)}{\sqrt{\alpha}} - 1\right)\bar r - \frac{1}{2\epsilon}\left(\frac{f'(u^\epsilon)}{\sqrt{\alpha}} + 1\right)\bar s. \end{cases} \tag{3.20}$$

Integrate the first equation against $sgn(\bar r)$, the second against $sgn(\bar s)$, and add; in view of the subcharacteristic condition (2.2) we find (compare [8, equation (4.10)])

$$\|\bar r\|_{L^1} + \|\bar s\|_{L^1} \le \|\bar r_0\|_{L^1} + \|\bar s_0\|_{L^1}. \tag{3.21}$$

If the initial data are prepared in the sense that $\|v_0^\epsilon - f(u_0^\epsilon)\|_{L^1} = \mathcal{O}(\epsilon)$ (and in fact, in our case we ignore initial errors by restricting attention to (1.2)), then initial time derivatives

$$\|v_t^\epsilon(\cdot, t=0)\|_{L^1} + \|u_t^\epsilon(\cdot, t=0)\|_{L^1}$$

are bounded, and by (3.21), they remain bounded in later time. In particular, $\|v_t^\epsilon(\cdot,t)\|_{L^1} \le \text{Const}$. This, together with the BV bound of u^ϵ (which follows from the Lip^+ stability), imply that the Lip'-size of the local truncation error is of order ϵ

$$\|\epsilon(v_t^\epsilon + \alpha u_x^\epsilon)_x\|_{Lip'} \le \epsilon(\|v_t^\epsilon\|_{L^1} + \alpha\|u_x^\epsilon\|_{L^1}) \le \mathcal{O}(\epsilon), \tag{3.22}$$

and consequently, (3.17) implies that the Lip' size of the global error, $u^\epsilon - u$, is of the same order of $\mathcal{O}(\epsilon)$. If we interpolate between this Lip' bound and the BV boundedness of $u^\epsilon - u$, we arrive at an L^1 convergence rate estimate of order $\mathcal{O}(\sqrt{\epsilon})$,

$$\|u^\epsilon - u\|_{L^1} \le \text{Const}\|u^\epsilon - u\|_{Lip'}^{1/2} \cdot \|u^\epsilon - u\|_{BV}^{1/2} \le \text{Const}\sqrt{\epsilon}.$$

The Lip^+ stability of u^ϵ enables us to convert this global estimate into a local one: using Lemma 2.3 with $(z,w) = (u^\epsilon, u)$ we find (see (2.10))

$$|u^\epsilon(x,t) - u(x,t)| \le \text{Const}\,\delta \cdot |u|_{C^1_{loc}(x-\delta,x+\delta)}, \quad \delta \sim \epsilon^{1/4}. \tag{3.23}$$

There are several possible improvements.

- If one utilizes the $\mathcal{O}(\epsilon)$-Lip' error estimate (instead of the L^1 estimate of order $\mathcal{O}(\sqrt{\epsilon})$), then this pointwise error estimate can be further improved outside a smaller shock region of size $\delta \sim \epsilon^{1/3}$ (see [14, Corollary 2.4]).
- Moreover, for piecewise smooth data one has an L^1-error estimate of order ϵ, [24], and the above arguments yield pointwise error estimate of order $\delta \sim \|u^\epsilon - u\|_{L^1}^{1/2} = \mathcal{O}(\sqrt{\epsilon})$; this will be outlined in section 3.3.
- Finally, in section 4 we will present a bootstrap argument for a further improvement of this pointwise error estimate; we prove an pointwise error of order δ outside a shock zone of optimal size $\delta \sim \epsilon$.

Remark. In (1.3) we restrict our attention to initial data which are *exactly* matched with their assumed limit, $v_0^\epsilon = f(u_0^\epsilon)$. It is clear from the above discussion that Lip' error bound of order $\mathcal{O}(\epsilon)$ holds for more general initial data, which are only required to be prepared so that $\|u_0^\epsilon - u_0\|_{Lip'} + \|v_0^\epsilon - f(u_0^\epsilon)\|_{L^1} = \mathcal{O}(\epsilon)$.

3.3. A nonoptimal pointwise error estimate. In the following section, we will consider the case that the entropy solution for (1.3) is piecewise smooth, with finitely many shock discontinuities. Thus, if we let $S(t)$ denote the singular support of $u(\bullet, t)$, then it consists of finitely many shocks, $S(t) := \{(x,t) \,|\, x = X_k(t)\}$, each of which satisfies the Rankine–Hugoniot and the Lax conditions:

$$X_k' = \frac{[f(u(X_k,t)]}{[u(X_k,t)]}, \tag{3.24}$$

$$f'(u(X_k(t)-,t)) > X_k'(t) > f'(u(X_k(t)+,t)). \tag{3.25}$$

We note in passing that many practical initial data lead to finite number of shocks (see, e.g., [17, 22]), and in this case one has a global L^1-error bound of order ϵ, (2.7). Next we consider the characteristic variables, $\sqrt{\alpha}u^\epsilon \pm v^\epsilon$: It follows that their L^1 convergence rate from their limiting value $\sqrt{\alpha}u \pm v$ with $v = f(u)$ is also order (ϵ). Moreover, Theorem 3.1 implies the Lip^+ boundedness

$$\sqrt{\alpha}u_x^\epsilon + v_x^\epsilon \le C, \qquad \sqrt{\alpha}u_x^\epsilon - v_x^\epsilon \le C.$$

We can now apply the interpolation inequality (2.10), with $(z,w) = (\sqrt{\alpha}u^\epsilon \pm v^\epsilon, \sqrt{\alpha}u \pm f(u))$. We obtain the following pointwise error bound (see also [20]):

$$\begin{cases} |\sqrt{\alpha}u^\epsilon + v^\epsilon - (\sqrt{\alpha}u + f(u))| \le C\sqrt{\epsilon}, \\ \\ |\sqrt{\alpha}u^\epsilon - v^\epsilon - (\sqrt{\alpha}u - f(u))| \le C\sqrt{\epsilon} \quad \text{for} \quad \operatorname{dist}(x, S(t)) \ge \sqrt{\epsilon}. \end{cases} \tag{3.26}$$

It follows from the above results that

$$\begin{cases} |u^\epsilon(x,t) - u(x,t)| \le C\sqrt{\epsilon}, \\ \\ |v^\epsilon(x,t) - f(u(x,t))| \le C\sqrt{\epsilon} \quad \text{for} \quad \operatorname{dist}(x, S(t)) \ge \sqrt{\epsilon}. \end{cases} \tag{3.27}$$

Although the above pointwise local estimate is not optimal, it will suffice to derive the optimal error bound by a bootstrap argument which employs the comparison Lemma 2.4.

4. Pointwise error estimate. The key tool in obtaining the optimal pointwise error estimate is to use Lemma 2.4. In order to use it, we need to construct appropriate functions u_1 and u_2 (in this section they are error functions) such that they satisfy (2.11) and those conditions listed in the lemma. To illustrate the main idea of our proof, we first concentrate on the case that there is only one shock curve.

4.1. The case of a single shock. We assume that there is a smooth curve, $S(t) := \{(x,t)\,|\,x = X(t)\}$, so that $u(x,t)$ is C^2-smooth at any point $x \neq X(t)$. There are two smooth regions $x > X(t)$ and $x < X(t)$. We consider the pointwise error estimate in the region $x > X(t)$; the results for $x < X(t)$ can be obtained in a similar way. The function $\phi(x) \in C^2([0,\infty))$ satisfies

$$\phi(x) \sim \begin{cases} x & \text{if} \quad 0 \le x \ll 1, \\ 1 & \text{if} \quad x \gg 1. \end{cases}$$

More precisely, the function ϕ satisfies

$$\begin{cases} \phi(0) = 0, \qquad \phi'(x) > 0, \quad \phi(x) \le x \quad \text{for} \quad x > 0; \\ x\phi'(x) \le \phi(x) \quad \text{for} \quad x \ge 0; \\ |\phi^{(k)}(x)| \le 1, \quad x \ge 0, \end{cases} \tag{4.1}$$

e.g., $\phi(x) = 1 - e^{-x}$. Roughly speaking, the weighted function behaves like $\phi(x) \sim \min(|x|, 1)$.

We define two functions, which roughly speaking are the errors for u^ϵ and v^ϵ, in the following form:

$$U = u^\epsilon - u - \sigma(x,t), \qquad V = v^\epsilon - f(u+\sigma) + \epsilon\Psi(x,t), \tag{4.2}$$

where

$$\begin{cases} \sigma = \epsilon d e^{\gamma t}/\phi(x - X(t)), \\ \Psi = \Big(\alpha - f'(u+\sigma)f'(u)\Big)u_x - \Big(\sqrt{\alpha} + f'(u+\sigma)\Big)\Big(\sqrt{\alpha} - \dot{X}(t)\Big)\sigma\phi'/\phi. \end{cases} \tag{4.3}$$

In the above definitions, $\phi = \phi(x - X(t))$ is the so-called weighted distance to the shock set.[2] Also, in the above definitions, d and γ are two positive numbers to be determined.

Remark. It is seen from (4.2) and (4.3) that U is the error function for u^ϵ with first-order correction $O(\epsilon)$, while V is the error function for v^ϵ with first- and second-order corrections.

4.1.1. The basic idea. In order to put the error functions U and V to the framework of Lemma 2.4, we further let

$$p = \sqrt{\alpha}U + V, \qquad q = \sqrt{\alpha}U - V \tag{4.4}$$

and will verify the following estimates:

- (C1): for $x \ge X(0) + \sqrt{\epsilon}$,

$$p(x,0) \le 0, \qquad q(x,0) \le 0.$$

- (C2): for all $t \ge 0$,

$$p(X(t) + \sqrt{\epsilon}, t) \le 0, \qquad q(X(t) + \sqrt{\epsilon}, t) \le 0.$$

[2] In the case $x < X(t)$, the weighted distance is $\phi(X(t) - x)$. In other words, the weighted distance for any choice of x is $\phi(|x - X(t)|)$ in the single shock case.

- (C3): The functions p and q satisfies the following equations:

$$\begin{cases} p_t + \sqrt{\alpha} p_x = \alpha_{11} p + \alpha_{12} q + \beta_1(x,t), \\ \\ q_t - \sqrt{\alpha} q_x = \alpha_{21} p + \alpha_{22} q + \beta_2(x,t), \end{cases} \tag{4.5}$$

  where for $x \geq X(t) + \sqrt{\epsilon}$ the coefficients α_{12} and α_{21} are nonnegative, and the source terms β_1 and β_2 are nonpositive.

The idea is to choose d and γ sufficiently large so that Lemma 2.4 can be applied. The estimates (C1) and (C2) are satisfied by choosing sufficiently large d. Then for the time interval $0 < t \leq T_1 := \gamma^{-1}$, i.e.,

$$e^{\gamma t} \leq e, \tag{4.6}$$

we show that (C3) is satisfied by choosing sufficiently large γ.

After showing that (C1)–(C3) are verified for $t \in [0, T_1]$, we know that the error bounds for u^ϵ and v^ϵ can be established for $0 \leq t \leq T_1$. We can then use $u^\epsilon(x, T_1)$ and $u(x, T_1)$ as new initial data and repeat the same procedure to obtain the local error bounds for $T_1 < t \leq T_2$. By this standard continuation arguments, we can obtain the error bounds up to $t = T$.

4.1.2. The verification for (C1). Observe that for $x \geq X(t) + \sqrt{\epsilon}$

$$|\Psi(x,t)| \leq C + C\epsilon^{-1/2}\sigma, \qquad \sigma \geq \epsilon d. \tag{4.7}$$

Since $u^\epsilon(x,0) = u(x,0) = u_0(x)$ and $v^\epsilon(x,0) = f(u_0)$, we have, for $x \geq X(0)+ \geq \sqrt{\epsilon}$,

$$\begin{aligned} p(x,0) &= -\sqrt{\alpha}\sigma(x,0) + f(u_0) - f(u_0 + \sigma) + \epsilon\Psi \\ &= \Big(-\sqrt{\alpha} + f'(\bullet)\Big)\sigma(x,0) + \epsilon\,\Psi \\ &\leq -C_1\sigma + C\epsilon + C\sqrt{\epsilon}\sigma \\ &\leq \Big(-C_1 + C/d + C\sqrt{\epsilon}\Big)\,\sigma \leq 0 \end{aligned}$$

provided that d is sufficiently large and ϵ sufficiently small. Similarly, we can show that $q(x,0) < 0$ for $x > X(0) + \sqrt{\epsilon}$ with sufficiently large d and small ϵ.

4.1.3. The verification for (C2). It follows from the nonoptimized local error estimates (3.27) that

$$u^\epsilon - u = \mathcal{O}(\sqrt{\epsilon}), \qquad v^\epsilon - f(u) = \mathcal{O}(\sqrt{\epsilon}) \quad \text{for} \quad x \geq X(t) + \sqrt{\epsilon}.$$

It is also observed that $\sigma(x,t) \geq Cd\sqrt{\epsilon}$ for $x = X(t) + \sqrt{\epsilon}$. From the definition of p we obtain

$$\begin{aligned} p(X(t) + \sqrt{\epsilon}, t) &= \sqrt{\alpha} \times \mathcal{O}(\sqrt{\epsilon}) - \sqrt{\alpha}\sigma(x,t) + \mathcal{O}(\sqrt{\epsilon}) + f(u) - f(u+\sigma) + \mathcal{O}(\epsilon) + \mathcal{O}(\epsilon\sigma) \\ &= \mathcal{O}(\sqrt{\epsilon}) + (-\sqrt{\alpha} - f'(\bullet))\sigma(X(t) + \sqrt{\epsilon}, t) + \mathcal{O}(\epsilon) + \mathcal{O}(\epsilon\sigma) \\ &\leq C\sqrt{\epsilon} - C_1\, C\sqrt{\epsilon}\, d + C\epsilon \leq 0 \end{aligned}$$

provided that d is sufficiently large. Similarly, we can show that $q(X(t) + \sqrt{\epsilon}, t) \leq 0$.

4.1.4. The verification for (C3). By the definitions of U and V, as well as the relaxation equations (1.1) and its equilibrium equation (1.3), we have

$$
\begin{aligned}
U_t + V_x &= \underbrace{u_t^\epsilon + v_x^\epsilon}_{=0} - \underbrace{(u_t + f(u)_x)}_{=0} + (f(u)_x - f(u+\sigma)_x) - \sigma_t + \epsilon\Psi_x \\
&= \Big(f'(u) - f'(u+\sigma)\Big)u_x - f'(u+\sigma)\sigma_x - \sigma_t + \epsilon\Psi_x \\
&= -f''(\bullet)u_x\sigma - f'(u+\sigma)\sigma_x - \sigma_t + \epsilon\Psi_x.
\end{aligned} \tag{4.8}
$$

Similarly, we calculate $V_t + \alpha U_x$ and obtain

$$
\begin{aligned}
V_t + \alpha U_x &= v_t^\epsilon + \alpha u_x^\epsilon - f(u+\sigma)_t - \alpha u_x + \epsilon\Psi_t - \alpha\sigma_x \\
&= \frac{1}{\epsilon}\Big(f(u^\epsilon) - v^\epsilon\Big) - f'(u+\sigma)(u_t + \sigma_t) - \alpha u_x + \epsilon\Psi_t - \alpha\sigma_x \\
&= \frac{1}{\epsilon}\Big(f(u^\epsilon) - f(u+\sigma)\Big) - \frac{1}{\epsilon}\Big(v^\epsilon - f(u+\sigma) + \epsilon\Psi\Big) \\
&\qquad -f'(u+\sigma)\sigma_t + \epsilon\Psi_t - \alpha\sigma_x + \Psi + \Big(f'(u+\sigma)f'(u) - \alpha\Big)u_x \\
&= \frac{1}{\epsilon}f'(\bullet)U - \frac{1}{\epsilon}V - f'(u+\sigma)\sigma_t + \epsilon\Psi_t - \alpha\sigma_x \\
&\qquad -\Big(\sqrt{\alpha} + f'(u+\sigma)\Big)\Big(\sqrt{\alpha} - \dot{X}(t)\Big)\frac{\sigma\phi'}{\phi}.
\end{aligned} \tag{4.9}
$$

By the definition of p, $p = \sqrt{\alpha}U + V$, we obtain from (4.8) and (4.9) that p satisfies the first equation in (4.5) with

$$
\begin{cases}
\alpha_{11} = \frac{1}{2\epsilon}\Big(\frac{f'(\bullet)}{\sqrt{\alpha}} - 1\Big), \qquad \alpha_{12} = \frac{1}{2\epsilon}\Big(\frac{f'(\bullet)}{\sqrt{\alpha}} + 1\Big), \\
\beta_1(x,t) = -\sqrt{\alpha}f''(\bullet)u_x\sigma - \Big(\sqrt{\alpha} + f'(u+\sigma)\Big)(\sqrt{\alpha}\sigma_x + \sigma_t) \\
\qquad\qquad +\epsilon\Big(\sqrt{\alpha}\Psi_x + \Psi_t\Big) - \Big(\sqrt{\alpha} + f'(u+\sigma)\Big)(\sqrt{\alpha} - \dot{X}(t))\frac{\sigma\phi'}{\phi}.
\end{cases} \tag{4.10}
$$

We observe that $\alpha_{12} \geq 0$. Now we need to verify that $\beta_1(x,t) \leq 0$ for $x \geq X(t) + \sqrt{\epsilon}$ provided that γ is sufficiently large. It follows from the definitions of σ that

$$
\sigma_x = -\frac{\sigma\phi'}{\phi}, \quad \sigma_t = \gamma\sigma + \frac{\sigma\phi'}{\phi}\dot{X}. \tag{4.11}
$$

Using the above results and the definition of Ψ gives

$$
\begin{aligned}
\Psi_x &= \mathcal{O}(1) + \mathcal{O}\left(\frac{\sigma^2}{\phi^2}\right) + \mathcal{O}\left(\frac{\sigma}{\phi^2}\right) \\
&\leq C + C\epsilon^{-1}\sigma^2 + C\epsilon^{-1}\sigma \leq C + C\epsilon^{-1}\sigma \qquad \Big(\text{using (4.6)}\Big); \\
\Psi_t &= \mathcal{O}(1) + \mathcal{O}\left(\frac{\sigma^2}{\phi^2}\right) + \mathcal{O}\left(\frac{\sigma}{\phi^2}\right) + \mathcal{O}\left(\frac{\sigma^2}{\phi}\right) \\
&\leq C + C\epsilon^{-1}\sigma + C\sigma.
\end{aligned}
$$

The above estimates, together with (4.11) and the definition of $\beta_1(x,t)$, yield

$$
\begin{aligned}
\beta_1(x,t) &= -\sqrt{\alpha}f''(\bullet)u_x\sigma - \Big(\sqrt{\alpha} + f'(u+\sigma)\Big)\gamma\sigma + \epsilon\Big(\sqrt{\alpha}\Psi_x + \Psi_t\Big) \\
&\leq C\sigma - C_1\gamma\sigma + C\epsilon.
\end{aligned} \tag{4.12}
$$

From the definition of ψ we have $\epsilon = \phi e^{-\gamma t}\sigma/d \le C\sigma/d$. This, together with (4.12), gives

$$\beta_1(x,t) \le \Big(C - C_1\gamma + C/d\Big)\sigma \le 0 \qquad \text{for} \quad x \ge X(t) + \sqrt{\epsilon} \tag{4.13}$$

provided that γ is sufficiently large.

It follows from (4.8)–(4.9) and the definition of q, $q\sqrt{\alpha}U - V$, that q satisfies the second equation in (4.5) with

$$\begin{cases} \alpha_{21} = \frac{1}{2\epsilon}\left(1 - \frac{f'(\bullet)}{\sqrt{\alpha}}\right), \qquad \alpha_{22} = -\frac{1}{2\epsilon}\left(1 + \frac{f'(\bullet)}{\sqrt{\alpha}}\right), \\ \beta_2(x,t) = -\sqrt{\alpha}f''(\bullet)u_x\sigma + \Big(\sqrt{\alpha} - f'(u+\sigma)\Big)\Big(\sqrt{\alpha}\sigma_x - \sigma_t\Big) + \epsilon\Big(\sqrt{\alpha}\Psi_x - \Psi_t\Big) \\ \qquad\qquad + \Big(\sqrt{\alpha} + f'(u+\sigma)\Big)(\sqrt{\alpha} - \dot{X})\frac{\sigma\phi'}{\phi}. \end{cases} \tag{4.14}$$

It is seen that $\alpha_{21} \ge 0$. Moreover, it follows from (4.11) that

$$\beta_2(x,t) = -\sqrt{\alpha}f''(\bullet)u_x\sigma - \Big(\sqrt{\alpha} - f'(u+\sigma)\Big)\gamma\sigma + \epsilon\Big(\sqrt{\alpha}\Psi_x + \Psi_t\Big) + J, \tag{4.15}$$

where the last term J is defined by

$$J = 2\sqrt{\alpha}\Big(f'(u+\sigma) - \dot{X}\Big)\frac{\sigma\phi'}{\phi}. \tag{4.16}$$

Let $u_+ := u(X(t)+0,t)$. Using Lax geometrical entropy condition $\dot{X}(t) \ge f'(u_+)$ gives

$$\begin{aligned} &f'(u+\sigma) - \dot{X}(t) \le f'(u+\sigma) - f'(u_+) \\ &= f''(\bullet)(u(x,t) - u(X(t)+0,t)) + f''(\bullet)\sigma \\ &= f''(\bullet)u_x(\bullet,t)(x - X(t)) + f''(\bullet)\sigma. \end{aligned}$$

Using the fact that $x\phi'(x) \le \phi(x)$ gives

$$(x - X(t))\phi'(x - X(t)) \le C\,\phi(x - X(t)).$$

By the definition of J and the above observations, we have

$$J \le C\sigma + C\frac{\sigma^2}{\phi} \le C\sigma,$$

where in the last step we have used the fact $\sigma/\phi \le Cd$. It follows from the above results and the equation for β_2, (4.15), that $\beta_2 \le 0$ provided that γ is sufficiently large.

In summary, if d and $\gamma = \gamma(d)$ are sufficiently large, then the comparison lemma, Lemma 2.4, gives

$$p(x,t) \le 0, \qquad q(x,t) \le 0 \qquad \text{for} \quad x \ge X(t) + \sqrt{\epsilon}. \tag{4.17}$$

Similarly, changing $\phi(x - X(t))$ in (4.3) to $\phi(X(t) - x)$ will handle the case for $x \le X(t) - \sqrt{\epsilon}$. We will then obtain the following results:

$$p(x,t) \le 0, \qquad q(x,t) \le 0 \qquad \text{for} \quad x \le X(t) - \sqrt{\epsilon}. \tag{4.18}$$

Since

$$u^\epsilon - u - \epsilon d e^{\gamma t}/\phi = \frac{1}{2\sqrt{\alpha}}(p+q),$$

the estimates (4.17) and (4.18) yield

$$u^\epsilon - u \le \epsilon\, d e^{\gamma t}/\phi \qquad \text{for} \quad |x - X(t)| \ge \sqrt{\epsilon}. \tag{4.19}$$

By letting $U = u - u^\epsilon - \sigma$ and $V = f(u-\sigma) - v^\epsilon + \epsilon\tilde{\Psi}$, where $\tilde{\Psi}$ is of a similar form for Ψ, we can again using the comparison lemma obtain

$$u - u^\epsilon \le \epsilon\, d e^{\gamma t}/\phi \qquad \text{for} \quad |x - X(t)| \ge \sqrt{\epsilon}. \tag{4.20}$$

We summarize what we have shown by stating the following.

ASSERTION 4.1. *Let $u^\epsilon(x,t)$ be the relaxation solutions of* (1.1)–(1.2) *and $u(x,t)$ be the entropy solution of* (1.3). *If the entropy solution has* only one *shock discontinuity $S(t) = \{(x,t)|x = X(t)\}$, then the following error estimates hold:*

- *For a weighted distance function ϕ, $\phi(x) \sim \min(|x|, 1)$,*

$$|(u^\epsilon - u)(x,t)|\phi(|x - X(t)|) = \mathcal{O}(\epsilon), \quad |x - X(t)| \ge \sqrt{\epsilon}. \tag{4.21}$$

- *In particular, if (x,t) is away from the singular support, then*

$$|(u^\epsilon - u)(x,t)| \le C(h)\epsilon, \qquad \mathrm{dist}(x, S(t)) \ge h. \tag{4.22}$$ ∎

4.2. Finitely many shocks. In the case that the entropy solutions for the conservation law (1.3) have two shocks, our analysis in section 4.1 can be extended to cover this case easily. The main difference is to change the weighted distance function ϕ to the product of two weighted distance functions, i.e., $\phi(|x - X_1(t)|) \cdot \phi(|x - X_2(t)|)$. This idea was used in [20].

In a more general case when there are finitely many shocks, we replace the weighted distance function by

$$\rho(x,t) = \prod_{k=1}^{K} \phi\Big(|x - X_k(t)|\Big). \tag{4.23}$$

Then we consider the error functions similar to (4.3). We can apply the same techniques as used in the last subsection to obtain the optimal error bounds. We omit the detail procedure but state our main result as follows.

THEOREM 4.1. *Let $u^\epsilon(x,t)$ be the relaxation solutions of* (1.1)–(1.2) *and $u(x,t)$ be the entropy solution of* (1.3). *If the entropy solution has* finitely *many shock discontinuities, $S(t) = \{(x,t)|x = X_k(t)\}_{k=1}^K$, then the following error estimates hold:*

- *For a weighted distance function ϕ, $\phi(x) \sim \min(|x|, 1)$,*

$$|(u^\epsilon - u)(x,t)| \prod_{k=1}^{K} \phi\Big(|x - X_k(t)|\Big) = \mathcal{O}(\epsilon), \quad |x - X(t)| \ge \sqrt{\epsilon}. \tag{4.24}$$

- *In particular, if (x,t) is away from the singular support, then*

$$|(u^\epsilon - u)(x,t)| \le C(h)\epsilon, \qquad \mathrm{dist}(x, S(t)) \ge h. \tag{4.25}$$ ∎

It is noted that the above results cover the case when there are finitely many shocks with *possible collisions*.

5. Concluding remarks. In this work, we have obtained the pointwise error bounds for relaxation approximations to scalar conservation laws with piecewise smooth solutions. The proof of our results is based on two ingredients: a one-sided interpolation inequality (interpolating the L^1 error estimates and Lip^+ stability bounds), and a comparison theorem for weakly coupled hyperbolic systems. Here, we only investigated the case of entropy solutions of the equilibrium equation (1.3) which consist of finitely many shocks. The techniques used in this paper can be extended, however, in several directions:

- *Finitely many rarefaction waves.* Combining the techniques presented in [20], sharp pointwise error bounds can be obtained for entropy solutions of the equilibrium equation (1.3) which consist of finitely many rarefactions.
- *Finite difference approximations.* Sharp pointwise error bounds can be obtained for difference approximations of the equilibrium equation (1.3). A convergence study based on Lip' arguments was presented in [14]. Augmented with one-sided interpolation together with appropriate comparison techniques along the lines of our discussion in section 4, one can convert the global Lip' error estimates into sharp pointwise error estimates. The example of Lax–Friedrichs central scheme, corresponding to the first-order relaxation scheme of [6], was worked out by the authors in [21]. The second-order schemes based on the relaxation approximation (1.1) correspond to the central scheme in [13], and like most high-resolution schemes, the main difficulty lies with the question of their Lip^+ stability.

REFERENCES

[1] G.-Q. Chen, C. Levermore, and T. P. Liu, *Hyperbolic conservation laws with stiff relaxation terms and entropy*, Comm. Pure Appl. Math., 47 (1994), pp. 787–830.

[2] B. Engquist and B. Sjogreen, *The convergence rate of finite difference schemes in the presence of shocks*, SIAM J. Numer. Anal., 35 (1998), pp. 2464–2485.

[3] B. Engquist and S.-H. Yu, *Convergence of Finite Difference Schemes for Piecewise Smooth Solutions with Shocks*, preprint, 1997.

[4] J. Goodman and Z. Xin, *Viscous limits for piecewise smooth solutions to systems of conservation laws*, Arch. Rational Mech. Anal., 121 (1992), pp. 235–265.

[5] G.-S. Jiang and E. Tadmor, *Nonoscillatory central schemes for multidimensional hyperbolic conservation laws*, SIAM J. Sci. Comput., 19 (1998), pp. 1892–1917.

[6] S. Jin and Z. Xin, *The relaxation schemes for system of conservation laws in arbitrary space dimensions*, Comm. Pure Appl. Math., 48 (1995), pp. 235–277.

[7] M. A. Katsoulakis and A. E. Tzavaras, *Contractive relaxation systems and the scalar multidimensional conservation law*, Comm. Partial Differential Equations, 22 (1997), pp. 195–233.

[8] A. Kurganov and E. Tadmor, *Stiff systems of hyperbolic conservation laws: Convergence and error estimates*, SIAM J. Math. Anal., 28 (1997), pp. 1446–1456.

[9] N. N. Kuznetsov, *Accuracy of some approximate methods for computing the weak solutions of a first-order quasi-linear equation*, U.S.S.R. Comput. Math. and Math. Phys., 16 (1976), pp. 105–119.

[10] P. L. Lions, B. Perthame, and E. Tadmor, *Kinetic formulation of scalar conservation laws*, J. Amer. Math. Soc., 7 (1994), pp. 169–191.

[11] T. P. Liu, *Hyperbolic conservation laws with relaxation*, Comm. Math. Phys., 108 (1987), pp. 153–175.

[12] R. Natalini, *Convergence to equilibrium for the relaxation approximations of conservation laws*, Comm. Pure Appl. Math., 49 (1996), pp. 795–823.

[13] H. Nessyahu and E. Tadmor, *Non-oscillatory central differencing for hyperbolic conservation laws*, J. Comput. Phys., 87 (1990), pp. 408–463.

[14] H. Nessyahu and E. Tadmor, *The convergence rate of approximate solutions for nonlinear scalar conservation laws*, SIAM J. Numer. Anal., 29 (1992), pp. 1505–1519.

[15] B. Perthame and E. Tadmor, *A kinetic equation with kinetic entropy functions for scalar conservation laws*, Comm. Math. Phys., 136 (1991), pp. 501–517.
[16] M. H. Protter and H. F. Weinberger, *Maximum Principles in Differential Equations*, Prentice–Hall, Englewood Cliffs, NJ, 1967.
[17] D. G. Schaeffer, *A regularity theorem for conservation laws*, Adv. Math., 11 (1973), pp. 368–386.
[18] H. J. Schroll, A. Tveito, and R. Winther, *An L^1-error bound for a semi-implicit difference scheme applied to a stiff system of conservation laws*, SIAM J. Numer. Anal., 34 (1997), pp. 1152–1166.
[19] E. Tadmor, *Local error estimates for discontinuous solutions of nonlinear hyperbolic equations*, SIAM J. Numer. Anal., 28 (1991), pp. 891–906.
[20] E. Tadmor and T. Tang, *Pointwise error estimates for scalar conservation laws with piecewise smooth solutions*, SIAM J. Numer. Anal., 36 (1999), pp. 1739–1758.
[21] E. Tadmor and T. Tang, *Pointwise convergence rate for nonlinear conservation laws*, in Hyperbolic Problems: Theory, Numerics, Applications, M. Fey and R. Jeltsch, eds., Internat. Ser. Numer. Math. 130, Birkhäuser, Basel, 1999, pp. 925–934.
[22] E. Tadmor and T. Tassa, *On the piecewise smoothness of entropy solutions to scalar conservation laws*, Comm. Partial Differential Equations, 18 (1993), pp. 1631–1652.
[23] T. Tang and Z.-H. Teng, *Viscosity methods for piecewise smooth solutions to scalar conservation laws*, Math. Comp., 66 (1997), pp. 495–526.
[24] Z. H. Teng, *First-order L^1-convergence for relaxation approximations to conservation laws*, Comm. Pure Appl. Math., 51 (1998), pp. 857–895.
[25] Z.-H. Teng and P. Zhang, *Optimal L^1-rate of convergence for the viscosity method and monotone scheme to piecewise constant solutions with shocks*, SIAM J. Numer. Anal., 34 (1997), pp. 959–978.
[26] A. Tveito and R. Winther, *On the rate of convergence to equilibrium for a system of conservation laws including a relaxation term*, SIAM J. Math. Anal., 28 (1997), pp. 136–161.
[27] W.-C. Wang and Z. Xin, *Asymptotic limit of initial boundary value problems for conservation laws with relaxation extensions*, Comm. Pure Appl. Math., 51 (1998), pp. 505–535.
[28] G. Whitham, *Linear and Nonlinear Waves*, Wiley–Interscience, New York, 1974.

SIAM J. MATH. ANAL.
Vol. 32, No. 4, pp. 887–915

LONG TIME ASYMPTOTICS OF SOLUTIONS TO THE ANHARMONIC OSCILLATOR MODEL FROM NONLINEAR OPTICS*

FRANK JOCHMANN†

Abstract. The anharmonic oscillator model describing the propagation of electromagnetic waves in an exterior domain containing a nonlinear dielectric medium is investigated. The system under consideration consists of a generally nonlinear second order differential equation for the dielectrical polarization coupled with Maxwell's equations for the electromagnetic field. Local decay of the electromagnetic field for $t \to \infty$ in the charge free case is shown for a large class of potentials.

Key words. nonlinear optics, Maxwell's equations, exterior boundary value problem, asymptotic behavior

AMS subject classifications. 35Q60, 35L40, 78A35

PII. S0036141099360932

1. Introduction.

The subject of this paper is the anharmonic oscillator model from nonlinear optics consisting of Maxwell's equations

$$\partial_t \mathbf{E} = \operatorname{curl} \mathbf{H} - \partial_t \tilde{\mathbf{P}} - \mathbf{j}, \quad \partial_t \mathbf{H} = -\operatorname{curl} \mathbf{E} \tag{1.1}$$

on $\mathbb{R}^+ \times \Omega$ coupled with the equation

$$\alpha \partial_t^2 \mathbf{P} + \partial_t \mathbf{P} + \nabla_P V(x, \mathbf{P}) = \gamma \mathbf{E} \tag{1.2}$$

on $\mathbb{R}^+ \times G$. The initial boundary conditions

$$\vec{n} \wedge \mathbf{E} = 0 \text{ on } (0,\infty) \times \Gamma_1 \text{ and } \vec{n} \wedge \mathbf{H} = 0 \text{ on } (0,\infty) \times \Gamma_2, \tag{1.3}$$

$$\mathbf{E}(0,x) = \mathbf{E}_0(x), \mathbf{H}(0,x) = \mathbf{H}_0(x), \tag{1.4}$$

and

$$\mathbf{P}(0,x) = \mathbf{P}_0(x), \quad \partial_t \mathbf{P}(0,x) = \mathbf{P}_1(x) \text{ on } G \tag{1.5}$$

are imposed. This system describes the propagation of electromagnetic waves in a dielectric medium occupying the set G; see [3], [12]. Here $\Omega \subset \mathbb{R}^3$ is an exterior domain, $G \subset \Omega$ a certain subset, and $\Gamma_1 \subset \partial\Omega$, $\Gamma_2 \stackrel{\text{def}}{=} \partial\Omega \setminus \Gamma_1$. The unknown functions are the electric and magnetic field $\mathbf{E}, \mathbf{H}$, which depend on the time $t \geq 0$ and the space-variable $x \in \Omega$ and the dielectric polarization $\mathbf{P}$ defined on $\mathbb{R}^+ \times G$. In (1.1) the function $\tilde{\mathbf{P}}$ is the extension of $\mathbf{P}$ on $\mathbb{R}^+ \times \Omega$ defined by zero on the set $\mathbb{R}^+ \times (\Omega \setminus G)$. The physical meaning of the boundary condition (1.3) is that Γ_1 is perfectly conducting, such that the tangential component of the electric field must vanish.

*Received by the editors September 1, 1999; accepted for publication (in revised form) July 17, 2000; published electronically December 20, 2000. This work was supported by the Deutsche Forschungsgemeinschaft through SFB 555.

http://www.siam.org/journals/sima/32-4/36093.html

†Institut für Angewandte Mathematik, Sitz: Rudower Chaussee 25, Humboldt Universität Berlin, 10099 Berlin, Germany (jochmann@mathematik.hu-berlin.de).

The coefficients $\alpha, \gamma \in L^\infty(G)$ depending on the space variables take into account the possibly variable mass, electrical charge, and density of the oscillating charged particles. An external current $\mathbf{j} \in L^1((0,\infty), L^2(\Omega))$ is included also. The potential energy function $V : G \times \mathbb{R}^3 \to [0,\infty)$ causes a spring force $\nabla_P V(x, \mathbf{P})$, which may depend nonlinearly on $\mathbf{P}$. It is assumed that the potential V satisfies the attraction condition

$$0 \le V(x,y) \le Ky(\nabla_P V)(x,y) \text{ for all } x \in G, y \in \mathbb{R}^3 \tag{1.6}$$

with some constant $K > 0$.

In particular it is allowed that $|(\nabla_P V)(x,y)|$ tends to zero for $|y| \to \infty$ as in [12]. The linear case $(\nabla_P V)(x,y) = ay$ with some $a > 0$ is included also.

In [12], where $G = \Omega = \mathbb{R}^3$ and the coefficients and the potential do not depend on x, it is shown that (1.1)–(1.2) admits a unique strong solution in $C([0,\infty), H^s(\mathbb{R}^3))$ for $s \ge 2$. Note that in our case system (1.1) does not admit classical solutions on all of $(0,\infty) \times \Omega$ due to the discontinuity of $\tilde{\mathbf{P}}$ on $\Sigma \stackrel{\text{def}}{=} (\partial G) \cap \Omega$, the interface between the polarizable medium and the vacuum-region $\Omega \setminus G$. But if the solution is smooth on $(0,\infty) \times G$ and on $(0,\infty) \times (\Omega \setminus G)$, then (1.1) involves a transmission condition, which requires the continuity of the tangential components of $\mathbf{E}$ and $\mathbf{H}$, as well as a linking condition for the normal components of $\mathbf{D} = \mathbf{E} + \tilde{\mathbf{P}}$ and $\mathbf{H}$ on Σ. Therefore a suitable weak formulation of (1.1)–(1.2) will be given in section 2, which admits discontinuous solutions. In [4] the Landau–Lifschitz equation for the magnetic moment coupled with Maxwell's equations is handled analogously. The magnetic moment is located in a bounded domain, whereas Maxwell's equations are posed on the whole space. It is shown in [4] that all points of the weak ω-limit set are solutions of the corresponding stationary equations.

The main topic of this paper is the investigation of the long time asymptotic behavior of the solutions. For this purpose it is assumed that

$$\gamma \in L^{3/2}(G) \text{ and } (1+|x|)\gamma \in L^{r_0}(G) \text{ with some } r_0 \in (3/2, \infty).$$

Since $\gamma \in L^\infty(G)$, this assumption is fulfilled, for example, if $\int_G (1+|x|)^{r_0} dx < \infty$, in particular if the set G is bounded.

Let X_0 denote the set of all $(\mathbf{f}, \mathbf{g}) \in X \stackrel{\text{def}}{=} L^2(\Omega, \mathbb{R}^6)$ which satisfy

$$\text{curl } \mathbf{f} = \text{ curl } \mathbf{g} = 0 \text{ on } \Omega, \quad \vec{n} \wedge \mathbf{f} = 0 \text{ on } \Gamma_1, \vec{n} \wedge \mathbf{g} = 0 \text{ on } \Gamma_2.$$

The basic goal is to prove the decay property

$$\int_{\{x \in \Omega : |x| \le \alpha t\}} |\mathbf{E}|^2 + |\mathbf{H}|^2 dx \stackrel{t \to \infty}{\longrightarrow} 0 \text{ for all } \alpha < 1, \tag{1.7}$$

in particular local energy decay, provided that the initial data satisfy

$$\int_\Omega (\mathbf{D}_1 \mathbf{f} + \mathbf{H}_0 \mathbf{g}) dx = 0 \text{ for all } (\mathbf{f}, \mathbf{g}) \in X_0. \tag{1.8}$$

Here $\mathbf{D}_1 \stackrel{\text{def}}{=} \mathbf{E}_0 + \tilde{\mathbf{P}}_0 - \int_0^\infty \mathbf{j}(s) ds$, where $\tilde{\mathbf{P}}_0$ denotes the extension of $\mathbf{P}_0$ by zero on $\Omega \setminus G$. (Note that the propagation speed of electromagnetic waves in a vacuum is normalized to 1 in (1.1).) Furthermore it is shown that

$$\int_\Omega |\mathbf{E}(t,x) + |x|^{-1} x \wedge \mathbf{H}(t,x)|^2 + |\mathbf{H}(t,x) - |x|^{-1} x \wedge \mathbf{E}(t,x)|^2 dx \stackrel{t \to \infty}{\longrightarrow} 0. \tag{1.9}$$

The physical meaning of (1.7) is that the wave-packet $(\mathbf{E}(t), \mathbf{H}(t))$ is concentrated near the sphere $|x| = t$ for large times. In section 4 it is also shown that the solution $(\mathbf{E}(t), \mathbf{H}(t))$ behaves asymptotically like a solution of the linear homogeneous Maxwell equations in $\mathbb{R}^3$ as $t \to \infty$.

Condition (1.8) includes

$$\text{div } \mathbf{D}_1 = 0 \text{ and } \text{ div } \mathbf{H}_0 = 0 \text{ on } \Omega$$

and the boundary conditions

$$\vec{n}\mathbf{D}_1 = 0 \text{ on } \Gamma_2 \text{ and } \vec{n}\mathbf{H}_0 = 0 \text{ on } \Gamma_1.$$

By (1.1) the function $\mathbf{D} \stackrel{\text{def}}{=} \mathbf{E} + \tilde{\mathbf{P}}$ and $\mathbf{H}$ obey $\text{ div } \mathbf{H}(t) = \text{ div } \mathbf{H}_0 = 0$ and

$$\text{div } \mathbf{D}(t) = \text{ div } \left[\mathbf{E}_0 + \tilde{\mathbf{P}}_0 - \int_0^t \mathbf{j}(s)ds\right] \stackrel{t\to\infty}{\longrightarrow} \text{ div } \mathbf{D}_1 = 0 \text{ in } \mathcal{D}'(\Omega)$$

if condition (1.8) is fulfilled. Physically this means that the space charge $\rho \stackrel{\text{def}}{=} \text{ div } \mathbf{D}$ determined by the initial-state $(\mathbf{E}_0, \mathbf{H}_0)$ and the prescribed current $\mathbf{j}$ vanishes as $t \to \infty$.

The proof of the decay property (1.7) uses a result in [11]. In particular it is shown in section 3 that for arbitrary initial states $(\mathbf{E}_0, \mathbf{H}_0)$, $\mathbf{P}_0$ and $\mathbf{P}_1$ not necessarily satisfying (1.8), the weak ω-limit set of $(\mathbf{E}, \mathbf{H})$ is contained in X_0.

2. Basic definitions, assumptions, and preliminaries. For an arbitrary open set $K \subset \mathbb{R}^3$ the space of all infinitely differentiable functions with compact support contained in K is denoted by $C_0^\infty(K)$. For $p \in [1, \infty)$ the dual exponent p^* is given by $p^{-1} + (p^*)^{-1} = 1$.

Let $\Omega \subset \mathbb{R}^3$ be a (connected) domain with bounded complement such that $\mathbb{R}^3 \backslash \overline{\Omega}$ is a Lipschitz domain and $G \subset \Omega$ a measurable set with nonempty interior. Throughout this paper the following assumptions are imposed on $V : G \times \mathbb{R}^3 \to [0, \infty)$. First $V(\cdot, y) \in L^\infty(G)$ for all $y \in \mathbb{R}^3$,

$$V(x, \cdot) \in C^2(\mathbb{R}^3, \mathbb{R}), \quad V(x, 0) = 0, \text{ and } (\nabla_P V)(x, 0) = 0 \tag{2.1}$$

for all $x \in G$. It is assumed that $(\nabla_P V)$ is Lipschitz-continuous with respect to y, i.e., there exits some $L_0 \in (0, \infty)$, such that

$$|(\nabla_P V)(x, y) - (\nabla_P V)(x, z)| \leq L_0|y - z| \text{ for all } x \in G, y, z \in \mathbb{R}^3. \tag{2.2}$$

This condition is also required in [12], since the second and third order derivatives of V are assumed to be globally bounded there.

Next, let $\alpha \in L^\infty(G)$ be a uniformly positive and $\gamma \in L^\infty(G)$ be a positive, but not necessarily uniformly positive, function on G. Now $\mathcal{G} \subset L^2(G)$ is the weighted $L^2(G)$-space consisting of all measurable functions $\mathbf{f} : G \to \mathbb{R}^3$ with $\int_G \gamma^{-1}(x)|\mathbf{f}(x)|^2 dx < \infty$ endowed with the norm

$$\|\mathbf{f}\|_{\mathcal{G}}^2 \stackrel{\text{def}}{=} \int_G \gamma^{-1}(x)|\mathbf{f}(x)|^2 dx.$$

In what follows we denote by $\underline{\mathbf{w}}_1 \in \mathbb{C}^3$ the first three and by $\underline{\mathbf{w}}_2 \in \mathbb{C}^3$ the last three components of a vector $\mathbf{w} \in \mathbb{C}^6$ and $S\mathbf{w} \stackrel{\text{def}}{=} (-x \wedge \underline{\mathbf{w}}_2, x \wedge \underline{\mathbf{w}}_1)$.

Next, some function-spaces related to Maxwell's equations with mixed boundary conditions are introduced.

First W_H denotes the closure of $C_0^\infty(\mathbb{R}^3 \setminus \overline{\Gamma_2}, \mathbb{C}^3)$ in $H_{curl}(\Omega)$, where $H_{curl}(\Omega)$ is the space of all $\mathbf{E} \in L^2(\Omega, \mathbb{C}^3)$ with $\operatorname{curl} \mathbf{E} \in L^2(\Omega)$ in the sense of distributions.

Next, W_E denotes the set of all $\mathbf{E} \in H_{curl}(\Omega)$ such that

$$\int_\Omega \mathbf{E} \operatorname{curl} \mathbf{F} - \mathbf{F} \operatorname{curl} \mathbf{E} dx = 0 \text{ for all } \mathbf{F} \in W_H,$$

which includes a weak formulation of the boundary condition $\vec{n} \wedge \mathbf{E} = 0$ on Γ_1; see [7].

Now, the following operators are defined.

Let $D(B) \stackrel{\text{def}}{=} W_E \times W_H$ and

$$B(\mathbf{E}, \mathbf{H}) \stackrel{\text{def}}{=} (\operatorname{curl} \mathbf{H}, -\operatorname{curl} \mathbf{E}) \text{ for } (\mathbf{E}, \mathbf{H}) \in D(B).$$

Then B is a densely defined skew self-adjoint operator in the Hilbert-space $X \stackrel{\text{def}}{=} L^2(\Omega, \mathbb{C}^6)$ endowed with the usual scalar product. The space X_0 in (1.8) is defined as the kernel of B, i.e.,

$$X_0 \stackrel{\text{def}}{=} \{(\mathbf{E}, \mathbf{H}) \in D(B) : B(\mathbf{E}, \mathbf{H}) = 0\}$$

$$= \{(\mathbf{E}, \mathbf{H}) \in W_E \times W_H : \operatorname{curl} \mathbf{E} = \operatorname{curl} \mathbf{H} = 0\}.$$

Let Q be the orthogonal projector on $X_0^\perp = (\ker B)^\perp = \overline{\operatorname{ran} B}$.

For $\mathbf{f} \in L^1_{loc}([0, \infty), X)$ a function $\mathbf{u} \in C([0, \infty), X)$ is called a weak solution to the initial boundary value problem

$$\partial_t \underline{\mathbf{u}}_1 = \operatorname{curl} \underline{\mathbf{u}}_2 + \underline{\mathbf{f}}_1, \quad \partial_t \underline{\mathbf{u}}_2 = -\operatorname{curl} \underline{\mathbf{u}}_1 + \underline{\mathbf{f}}_2, \tag{2.3}$$

supplemented by the initial-boundary conditions

$$\vec{n} \wedge \underline{\mathbf{u}}_1 = 0 \text{ on } (0, \infty) \times \Gamma_1 \text{ and } \vec{n} \wedge \underline{\mathbf{u}}_2 = 0 \text{ on } (0, \infty) \times \Gamma_2 \tag{2.4}$$

if

$$\frac{d}{dt} \langle \mathbf{u}(t), \mathbf{a} \rangle_X = -\langle \mathbf{u}(t), B\mathbf{a} \rangle_X + \langle \mathbf{f}(t), \mathbf{a} \rangle_X \text{ for all } \mathbf{a} \in D(B). \tag{2.5}$$

This means that (2.3) is fulfilled in the sense of distributions, whereas the boundary conditions (2.4) are satisfied in the sense that $\int_0^t \mathbf{u}(s) ds \in D(B) = W_E \times W_H$ for all $t \geq 0$. It is well known that (2.5) is equivalent to the variation of constant formula

$$\mathbf{u}(t) = \exp(tB)\mathbf{u}(0) + \int_0^t \exp((t-s)B)\mathbf{f}(s) ds \tag{2.6}$$

where $(\exp(tB))_{t \in \mathbb{R}}$ is the unitary group generated by B; see [13]. (2.6) yields the energy estimate

$$\frac{1}{2} \frac{d}{dt} \|\mathbf{u}(t)\|_X^2 = \langle \mathbf{f}(t), \mathbf{u}(t) \rangle_X . \tag{2.7}$$

Next $\mathcal{R} : L^2(G) \to X$ is defined by

$$(\mathcal{R}\mathbf{p})(x) \stackrel{\text{def}}{=} (\mathbf{p}(x), 0) \text{ if } x \in G \text{ and } (\mathcal{R}\mathbf{p})(x) \stackrel{\text{def}}{=} 0 \text{ if } x \in \Omega \setminus G.$$

Let

$$\mathbf{j} \in L^1((0,\infty), L^2(\Omega, \mathbb{R}^3)), \quad (\mathbf{E}_0, \mathbf{H}_0) \in X, \quad \mathbf{P}_0 \in \mathcal{G} \text{ and } \mathbf{P}_1 \in \mathcal{G}. \tag{2.8}$$

By (2.1) and (2.2) the nonlinear composition operator $\mathbf{p} \in \mathcal{G} \to (\nabla_y V)(\cdot, \mathbf{p}(\cdot))$ is globally Lipschitz-continuous as a map from $\mathcal{G}$ to $\mathcal{G}$. Therefore the initial value problem

$$\alpha \partial_t^2 \mathbf{P} + \partial_t \mathbf{P} + (\nabla_y V)(x, \mathbf{P}) = \gamma \mathbf{E} \text{ on } (0,\infty) \times G \tag{2.9}$$

supplemented by the initial-conditions

$$\mathbf{P}(0) = \mathbf{P}_0, \quad \partial_t \mathbf{P}(0) = \mathbf{P}_1 \tag{2.10}$$

admits for all $\mathbf{E} \in C([0,\infty), L^2(\Omega, \mathbb{R}^3))$ a unique weak solution $\mathbf{P} \in C^2([0,\infty), \mathcal{G}) \subset C^2([0,\infty), L^2(G))$. If $\mathbf{E} \in C([0,\infty), L^2(\Omega, \mathbb{R}^3))$ and $\mathbf{F} \in C([0,\infty), L^2(\Omega, \mathbb{R}^3))$, then the Lipschitz continuity of $\nabla_y V$ yields the estimate

$$\frac{1}{2}\frac{d}{dt}\left[\|\alpha^{1/2}(\partial_t \mathbf{P}(t) - \partial_t \mathbf{Q}(t))\|_{\mathcal{G}}^2 + \|\mathbf{P}(t) - \mathbf{Q}(t)\|_{\mathcal{G}}^2\right] = \int_G \gamma^{-1}(\partial_t \mathbf{P} - \partial_t \mathbf{Q})$$

$$\times [\gamma \mathbf{E} - \partial_t \mathbf{P} - \nabla_P V(x, \mathbf{P}) - \gamma \mathbf{F} + \partial_t \mathbf{Q} + \nabla_P V(x, \mathbf{Q}) + \mathbf{P} - \mathbf{Q}]\, dx$$

$$\leq C_1 \|\alpha^{1/2}(\partial_t \mathbf{P}(t) - \partial_t \mathbf{Q}(t))\|_{\mathcal{G}} \left(\|\mathbf{E}(t) - \mathbf{F}(t)\|_{L^2(\Omega)} + \|\mathbf{P}(t) - \mathbf{Q}(t)\|_{\mathcal{G}}\right)$$

$$\leq C_2 \left[\|\alpha^{1/2}(\partial_t \mathbf{P}(t) - \partial_t \mathbf{Q}(t))\|_{\mathcal{G}}^2 + \|\mathbf{P}(t) - \mathbf{Q}(t)\|_{\mathcal{G}}^2 + \|\mathbf{E}(t) - \mathbf{F}(t)\|_{L^2(\Omega)}^2\right]$$

with constants C_1, C_2 independent of $\mathbf{E}, \mathbf{F}$, and t. Here $\mathbf{Q} \in C^2([0,\infty), \mathcal{G})$ is the solution of (2.9) and (2.10) with $\mathbf{E}$ replaced by $\mathbf{F}$. By Gronwall's lemma one obtains

$$\|\partial_t(\mathbf{P}(t) - \mathbf{Q}(t))\|_{L^2(G)} \leq \|\gamma^{1/2}\|_{L^\infty(G)} \|\partial_t(\mathbf{P}(t) - \mathbf{Q}(t))\|_{\mathcal{G}} \tag{2.11}$$

$$\leq C_3 \int_0^t \exp\left(L(t-s)\right) \|\gamma(\mathbf{E}(t) - \mathbf{F}(s))\|_{L^2(\Omega)} ds$$

with some $L, C_3 > 0$ independent of $\mathbf{E}$, $\mathbf{F}$, and t.

Let $\mathcal{A} : C([0,\infty), X) \to C([0,\infty), X)$ be defined by

$$\left(\mathcal{A}(\mathbf{E}, \mathbf{H})\right)(t) \stackrel{\text{def}}{=} \exp(tB)(\mathbf{E}_0, \mathbf{H}_0) - \int_0^t \exp\left((t-s)B\right) \left[\mathcal{R}\partial_t \mathbf{P}(s) + (\mathbf{j}(s), 0)\right] ds,$$

where $\mathbf{P}$ solves (2.9) and (2.10).

Now $(\mathbf{E}, \mathbf{H}) \in C([0,\infty), X)$ and $\mathbf{P} \in C^2([0,\infty), \mathcal{G})$ solve (1.1)–(1.5) (in the sense of (2.5)) if

$$(\mathbf{E}(t), \mathbf{H}(t)) = \exp(tB)(\mathbf{E}_0, \mathbf{H}_0) \tag{2.12}$$

$$-\int_0^t \exp\left((t-s)B\right)\left[\mathcal{R}\partial_t \mathbf{P}(s) + (\mathbf{j}(s),0)\right] ds,$$

i.e.,

$$\mathcal{A}(\mathbf{E},\mathbf{H}) = (\mathbf{E},\mathbf{H}), \tag{2.13}$$

and $\mathbf{P}$ solves (2.9) and (2.10). It follows from the estimates (2.7) and (2.11) and the contraction mapping principle in the space $C([0,T],X)$ with arbitrary large $T > 0$ that the fixed point problem (2.13) has a unique solution on each finite time interval $(0,T)$ and hence a unique global solution on $(0,\infty)$.

THEOREM 2.1. *Problem* (1.1)–(1.5) *has a unique weak solution* $(\mathbf{E},\mathbf{H},\mathbf{P})$ *with the properties* $(\mathbf{E},\mathbf{H}) \in C([0,\infty),X)$ *and* $\mathbf{P} \in C^2([0,\infty),\mathcal{G})$.

Further regularity of the solution can be obtained under the additional regularity assumption

$$(\mathbf{E}_0,\mathbf{H}_0) \in D(B) \text{ and } \mathbf{j} \in W^{1,1}((0,\infty),L^2(\Omega)). \tag{2.14}$$

Then $\mathcal{R}\partial_t \mathbf{P}(\cdot) + (\mathbf{j},0) \in W^{1,1}_{loc}([0,\infty),X)$. By the result in [13, Corollary 2.5, sect. 4.2] it follows that

$$(\mathbf{E},\mathbf{H}) \in C^1([0,\infty),X) \cap C([0,\infty),D(B)) \tag{2.15}$$

is a strong solution of

$$\partial_t(\mathbf{E}(t),\mathbf{H}(t)) = B(\mathbf{E}(t),\mathbf{H}(t)) - \mathcal{R}\partial_t \mathbf{P}(t) - (\mathbf{j}(t),0).$$

REMARK 1. *It follows from* (2.15) *that all partial derivatives occurring in* (1.1) *and* (1.2) *belong to the space* $L^\infty_{loc}([0,\infty),L^2(\Omega))$. *In this sense the solution is strong. As described in the introduction* $(\mathbf{E}(t),\mathbf{H}(t))$ *is not in* $H^1(\Omega)$, *in general due to the mixed boundary conditions and the possible discontinuity of the polarization. However, the divergence-free part of the electromagnetic field satisfies by equation* (2.15) *curl* $\underline{[Q(\mathbf{E}(\cdot),\mathbf{H}(\cdot))]}_k \in L^\infty_{loc}([0,\infty),L^2(\Omega))$ *and* div $\underline{[Q(\mathbf{E}(\cdot),\mathbf{H}(\cdot))]}_k = 0$ *for* $k \in \{1,2\}$. *Hence* $Q(\mathbf{E}(\cdot),\mathbf{H}(\cdot)) \in L^\infty_{loc}([0,\infty),H^1(\mathcal{U}))$ *for all subdomains* $\mathcal{U} \subset \Omega$ *which have positive distance to* $\partial\Omega$.

It follows from (2.12) and the energy estimate (2.7) that

$$\frac{1}{2}\frac{d}{dt}\|(\mathbf{E}(t),\mathbf{H}(t))\|_X^2 = -\langle \mathcal{R}\partial_t \mathbf{P}(t) + (\mathbf{j}(t),0)), (\mathbf{E}(t),\mathbf{H}(t))\rangle_X \tag{2.16}$$

$$= -\int_G \mathbf{E}\partial_t \mathbf{P} dx - \int_\Omega \mathbf{E}\mathbf{j} dx,$$

whereas (2.9) yields

$$\frac{d}{dt}\left(1/2\|\alpha^{1/2}\partial_t \mathbf{P}(t)\|_{\mathcal{G}}^2 + \int_G \gamma^{-1} V(x,\mathbf{P}) dx\right) \tag{2.17}$$

$$= -\int_G \gamma^{-1}|\partial_t \mathbf{P}|^2 dx + \int_G \mathbf{E}\partial_t \mathbf{P} dx.$$

By (2.16) and (2.17) one obtains the energy estimate

$$\frac{1}{2}\frac{d}{dt}\left(\|(\mathbf{E}(t),\mathbf{H}(t))\|_X^2+\|\alpha^{1/2}\partial_t\mathbf{P}(t)\|_{\mathcal{G}}^2+2\int_G\gamma^{-1}V(x,\mathbf{P})dx\right) \tag{2.18}$$

$$=-\int_G\gamma^{-1}|\,\partial\mathbf{P}|^2dx-\int_\Omega\mathbf{E}\mathbf{j}dx\le\|(\mathbf{E}(t),\mathbf{H}(t))\|_X\|\mathbf{j}(t)\|_{L^2(\Omega)}-\|\partial_t\mathbf{P}(t)\|_{\mathcal{G}}^2.$$

In the next lemma elementary properties of the solution are shown.

LEMMA 2.2. (i)

$$(\mathbf{E},\mathbf{H})\in L^\infty((0,\infty),X),\quad \partial_t\mathbf{P}\in L^\infty((0,\infty),\mathcal{G})\cap L^2((0,\infty),\mathcal{G})$$

and

$$\gamma^{-1}V(x,\mathbf{P}(\cdot))\in L^\infty((0,\infty),L^1(G)).$$

(ii) *If* (2.14) *holds, one has*

$$(\mathbf{E},\mathbf{H})\in W^{1,\infty}((0,\infty),X)\cap L^\infty((0,\infty),D(B)).$$

Proof. Let

$$\mathcal{E}_t\overset{\text{def}}{=}\left(\|(\mathbf{E}(t),\mathbf{H}(t))\|_X^2+\|\alpha^{1/2}\partial_t\mathbf{P}(t)\|_{\mathcal{G}}^2+2\int_G\gamma^{-1}V(x,\mathbf{P})dx\right). \tag{2.19}$$

By (2.18) one has

$$\frac{1}{2}\frac{d}{dt}\mathcal{E}_t\le\mathcal{E}_t(\mathbf{w})^{1/2}\|\mathbf{j}(t)\|_{L^2(\Omega)}-\|\partial_t\mathbf{P}(t)\|_{\mathcal{G}}^2.$$

Since $\|\mathbf{j}(\cdot)\|_{L^2(\Omega)}\in L^1(0,\infty)$, this inequality yields (i).

If (2.14) holds it follows from (2.15) that

$$\partial_t^2(\mathbf{E}(t),\mathbf{H}(t))=B\partial_t(\mathbf{E}(t),\mathbf{H}(t))-\mathcal{R}\partial_t^2\mathbf{P}(t)-\partial_t(\mathbf{j}(t),0) \tag{2.20}$$

is satisfied weakly in the sense of (2.5). With a similar estimate as before one obtains using the global boundedness of (D_P^2V) by (2.2)

$$\frac{1}{2}\frac{d}{dt}\left(\|\partial_t(\mathbf{E}(t),\mathbf{H}(t))\|_X^2+\|\alpha^{1/2}\partial_t^2\mathbf{P}(t)\|_{\mathcal{G}}^2\right) \tag{2.21}$$

$$=-\int_\Omega\partial_t\mathbf{E}\partial_t\mathbf{j}dx-\int_G\gamma^{-1}|\partial_t^2\mathbf{P}|^2dx-\int_G\gamma^{-1}\partial_t^2\mathbf{P}\cdot(D_P^2V)(x,\mathbf{P})\cdot\partial_t\mathbf{P}dx$$

$$\le\|\partial_t(\mathbf{E}(t),\mathbf{H}(t))\|_X\|\partial_t\mathbf{j}(t)\|_{L^2(\Omega)}+C_1\|\partial_t\mathbf{P}\|_{\mathcal{G}}^2-b_0/2\|\partial_t^2\mathbf{P}\|_{\mathcal{G}}^2.$$

With part (i) and $\|\partial_t\mathbf{j}(\cdot)\|_{L^2(\Omega)}\in L^1(0,\infty)$ it follows that

$$\partial_t(\mathbf{E},\mathbf{H})\in L^\infty((0,\infty),X). \tag{2.22}$$

By part (i) one has $\partial_t \mathbf{P} \in L^\infty((0,\infty),\mathcal{G}) \subset L^\infty((0,\infty),L^2(G))$, in particular $\mathcal{R}\partial_t\mathbf{P} \in L^\infty((0,\infty),X)$. Since also $(\mathbf{j},0) \in L^\infty((0,\infty),X)$ by (2.14) and $(\mathbf{E},\mathbf{H}) \in C^1([0,\infty),X) \cap C([0,\infty),D(B))$ solves

$$\partial_t(\mathbf{E}(t),\mathbf{H}(t)) = B(\mathbf{E}(t),\mathbf{H}(t)) - \mathcal{R}\partial_t\mathbf{P}(t) - (\mathbf{j}(t),0),$$

one obtains $(\mathbf{E},\mathbf{H}) \in L^\infty([0,\infty),D(B))$. This completes the proof of part (ii). ∎

By (2.18), (2.19), the previous lemma, and $\|\mathbf{j}(\cdot)\|_{L^2(\Omega)} \in L^1(0,\infty)$ one has $\frac{d}{dt}\mathcal{E}(t) \in L^1(0,\infty)$, which implies the existence of the limit

$$\mathcal{E}_\infty \stackrel{\text{def}}{=} \lim_{t\to\infty} \mathcal{E}(t) \tag{2.23}$$

$$= \lim_{t\to\infty} \left(\|(\mathbf{E}(t),\mathbf{H}(t))\|_X^2 + \|\alpha^{1/2}\partial_t\mathbf{P}(t)\|_{\mathcal{G}}^2 + 2\int_G \gamma^{-1} V(x,\mathbf{P})dx \right).$$

The physical meaning of $\mathcal{E}(t)$ is the total energy of the system, i.e., the sum of the potential and kinetic energy of the oscillating particles and the energy of the electromagnetic field. The dissipation term $-\|\partial_t\mathbf{P}(t)\|_{\mathcal{G}}^2 = -\int_G \gamma^{-1}|\partial_t\mathbf{P}|^2 dx$ in the energy estimate (2.18) describes the dielectric losses of the medium. This energy dissipation does not result from an electrical conductivity. It also occurs in insulating materials if they are exposed to a rapidly oscillating electric field.

3. A weak convergence property of the solutions. In what follows the additional regularity assumption (2.14) will be imposed on the data for convenience. The following "unique continuation" principle is proved in [11], which holds even for arbitrary, not necessarily bounded, spatial domains. As in [11] it will be used in the investigation of the weak ω-limit set of the solution of (1.1)–(1.5).

THEOREM 3.1. *Suppose that* $\mathbf{g} \in X$ *obeys*

$$\underline{(\exp(tB)\mathbf{g})}_1 = 0 \text{ on } G \text{ for all } t \in \mathbb{R}. \tag{3.1}$$

Then $\mathbf{g} \in \ker B$.

This is a generalization of the unique continuation principle for the scalar wave equation in bounded domains, which is used in [5], [6], and [15]; see also [2].

Theorem 3.1 says that each solution $(\mathbf{e},\mathbf{f}) \in C(\mathbb{R},L^2(\Omega,\mathbb{R}^{M+N}))$ of the evolution equation $\partial_t(\mathbf{e},\mathbf{f}) = B(\mathbf{e},\mathbf{f})$ with the property that $\mathbf{e}(t,x) = 0$ for all $t \in \mathbb{R}$ and $x \in G$ satisfies $(\mathbf{e}(0),\mathbf{f}(0)) \in \ker B$. In contrast to the unique continuation principle for bounded domains it is necessary to require the condition $\mathbf{e}(t,x) = 0$ on G for all $t \in \mathbb{R}$ and not only for positive times. The basic idea of the proof of Theorem 3.1 is to show that for each $f \in C_0^\infty(\mathbb{R}\setminus\{0\})$ the function $f(iB)\mathbf{g}$ is real-analytic and vanishes on G. This implies $f(iB)\mathbf{g} = 0$ for all $f \in C_0^\infty(\mathbb{R}\setminus\{0\})$ and hence $\mathbf{g} \in \ker B$. Here the operator $f(iB)$ can be defined by the spectral theorem, since iB is self-adjoint in $L^2(\Omega,\mathbb{C}^6)$. If $f \in C_0^\infty(\mathbb{R})$, then bounded operator $f(iB)$ has the representation

$$f(iB)\mathbf{u} = (2\pi)^{-1/2}\int_{\mathbb{R}} \hat{f}(t)\exp(-tB)\mathbf{u}dt \text{ for all } \mathbf{u} \in X. \tag{3.2}$$

Here $\hat{f}$ denotes the Fourier-transform of f.

In what follows let ω_0 denote the ω-limit-set of the trajectory $(\mathbf{E},\mathbf{H})$ with respect to the weak topology of X, i.e., the set of all $\mathbf{g} \in X$, such that there exists a sequence $t_n \stackrel{n\to\infty}{\longrightarrow} \infty$ with $(\mathbf{E}(t_n),\mathbf{H}(t_n)) \stackrel{n\to\infty}{\longrightarrow} \mathbf{g}$ in X weakly.

THEOREM 3.2. *$Q(\mathbf{E}(t),\mathbf{H}(t)) \overset{t\to\infty}{\longrightarrow} 0$ in X weakly.*

Proof. Suppose $\mathbf{g} \in \omega_0$ and $t_n \overset{n\to\infty}{\longrightarrow} \infty$ with

$$(\mathbf{E}(t_n),\mathbf{H}(t_n)) \overset{n\to\infty}{\longrightarrow} \mathbf{g} \quad \text{in } X \text{ weakly.} \tag{3.3}$$

Let $\mathbf{u}_n(t) \overset{\text{def}}{=} (\mathbf{E}(t_n+t),\mathbf{H}(t_n+t)) \in X$ and $\mathbf{f}_n(t) \overset{\text{def}}{=} (\nabla_P V)(x,\mathbf{P}(t_n+t)) \in \mathcal{G}$ for $n \in \mathbb{N}$. First, we have by (2.6)

$$\mathbf{u}_n(t) = \exp(tB)\mathbf{u}_n(0) - \int_{t_n}^{t_n+t} \exp((t_n+t-s)B)\,(\partial_s \mathcal{R}\mathbf{P}(s) + (\mathbf{j}(s),0))\,ds,$$

which implies by Lemma 2.2(i) that

$$\|\mathbf{u}_n(t) - \exp(tB)\mathbf{u}_n(0)\|_X \le \int_{t_n}^{t_n+t} \|\mathcal{R}\partial_s \mathbf{P}(s)\|_X + \|\mathbf{j}(s)\|_{L^2(\Omega)} ds \overset{n\to\infty}{\longrightarrow} 0$$

for all $t \in \mathbb{R}$ and hence by (3.3) with $\mathbf{u}_n(0) = (\mathbf{E}(t_n),\mathbf{H}(t_n))$

$$\mathbf{u}_n(t) \overset{n\to\infty}{\longrightarrow} \mathbf{u}_\infty(t) \overset{\text{def}}{=} \exp(tB)\mathbf{g} \text{ in } X \text{ weakly for all } t \in \mathbb{R}. \tag{3.4}$$

Lemma 2.2 yields

$$\|\mathbf{f}_n(t) - \mathbf{f}_n(0)\|_{\mathcal{G}} \le C_1 \int_{[t_n,t_n+t]} \|\partial_t \mathbf{P}(s)\|_{\mathcal{G}} ds \overset{n\to\infty}{\longrightarrow} 0 \text{ for all } t \in \mathbb{R}. \tag{3.5}$$

Suppose $T > 0$. Then (3.5) implies for all $\varphi \in C_0^\infty((-T,T),\mathcal{G})$

$$\lim_{n\to\infty} \int_{-T}^{T} \langle \mathbf{f}_n(t), \partial_t \varphi(t)\rangle_{\mathcal{G}} dt = \lim_{n\to\infty} \int_{-T}^{T} \langle \mathbf{f}_n(0), \partial_t \varphi(t)\rangle_{\mathcal{G}} dt = 0. \tag{3.6}$$

Using $\partial_t \mathbf{P} \in L^2((0,\infty),\mathcal{G})$ again one obtains from (2.9), (3.4), and (3.6) that

$$\begin{aligned} \int_{-T}^{T}\int_G \underline{(\mathbf{u}_\infty)}_1 \partial_t \varphi dx dt &= \lim_{n\to\infty} \int_{-T}^{T}\int_G \underline{(\mathbf{u}_n)}_1 \partial_t \varphi dx dt \\ &= \lim_{n\to\infty} \int_{-T}^{T} \langle \gamma \mathbf{E}(t_n+t), \partial_t \varphi(t)\rangle_{\mathcal{G}} dt \\ &= \lim_{n\to\infty} \int_{-T}^{T} \langle \alpha \partial_t^2 \mathbf{P}(t_n+t) + \partial_t \mathbf{P}(t_n+t) + \mathbf{f}_n(t), \partial_t \varphi(t)\rangle_{\mathcal{G}} dt = 0 \end{aligned} \tag{3.7}$$

for all $\varphi \in C_0^\infty((-T,T),\mathcal{G})$, in particular

$$\partial_t \underline{(\mathbf{u}_\infty)}_1(t,x) = 0 \text{ for all } x \in G, t \in (-T,T). \tag{3.8}$$

Since $T > 0$ is chosen arbitrarily, (3.8) holds for all $t \in \mathbb{R}$.

With $(\mathbf{E},\mathbf{H}) \in L^\infty((0,\infty),D(B))$ by Lemma 2.2(ii), it follows that $\mathbf{g} \in D(B)$ and

$$\underline{(\exp(tB)B\mathbf{g})}_1(x) = \partial_t \underline{(\mathbf{u}_\infty)}_1(t,x) = 0 \text{ for all } t \in \mathbb{R} \text{ and } x \in G$$

by (3.8). Invoking Theorem 3.1 one obtains $B\mathbf{g} \in \ker B$, and hence $\|B\mathbf{g}\|_X^2 = -\langle \mathbf{g}, B^2\mathbf{g}\rangle_X = 0$, whence $\mathbf{g} \in \ker B$. Hence

$$\omega_0 \subset \ker B. \tag{3.9}$$

Since $(\mathbf{E}(t),\mathbf{H}(t))$ is bounded in X as $t \to \infty$ by Lemma 2.2(i) and zero is the only possible accumulation point of $Q(\mathbf{E}(\cdot),\mathbf{H}(\cdot))$ by (3.9), the assertion follows. ∎

4. Decay of the electromagnetic field. For all $\mathbf{a} \in \ker B$ one has by (2.12)

$$\langle (\mathbf{E}(t), \mathbf{H}(t)), \mathbf{a} \rangle_X$$

$$= \left\langle \exp(tB)(\mathbf{E}_0, \mathbf{H}_0) - \int_0^t \exp((t-s)B) \left(\mathcal{R}\partial_s \mathbf{P}(s) + (\mathbf{j}(t), 0)\right) ds, \mathbf{a} \right\rangle_X$$

$$= \left\langle (\mathbf{E}_0, \mathbf{H}_0) + \mathcal{R}\mathbf{P}(0) - \mathcal{R}\mathbf{P}(t) - \int_0^t (\mathbf{j}(s), 0) ds, \mathbf{a} \right\rangle_X$$

and hence

$$(1-Q)\left((\mathbf{E}(t), \mathbf{H}(t)) + \mathcal{R}\mathbf{P}(t) + \int_0^t (\mathbf{j}(s), 0) ds - (\mathbf{E}_0, \mathbf{H}_0) - \mathcal{R}\mathbf{P}(0)\right) = 0. \tag{4.1}$$

Recall that $1-Q$ is the orthogonal projector on $X_0 = \ker B$.

Throughout this section it is assumed that the initial-state $(\mathbf{E}_0, \mathbf{H}_0) \in X$ satisfies

$$(\mathbf{D}_1, \mathbf{H}_0) = (\mathbf{E}_0, \mathbf{H}_0) + \mathcal{R}\mathbf{P}_0 - \int_0^\infty (\mathbf{j}(s), 0) ds \in X_0^\perp,$$

i.e.,

$$(1-Q)\left((\mathbf{E}_0, \mathbf{H}_0) + \mathcal{R}\mathbf{P}_0 - \int_0^\infty (\mathbf{j}(s), 0) ds\right) = 0. \tag{4.2}$$

This is condition (1.8) on the initial states. It follows from (4.1) and (4.2) that

$$(1-Q)(\mathbf{E}(t), \mathbf{H}(t)) = (1-Q)(\mathbf{J}(t) - \mathcal{R}\mathbf{P}(t)) \tag{4.3}$$

with $\mathbf{J}(t) \stackrel{\text{def}}{=} \int_t^\infty (\mathbf{j}(s), 0) ds$.

The main goal of this section is the proof of the decay property (1.7). The main steps are summarized now. By a standard energy estimate it follows that, roughly speaking, the asymptotic propagation speed of the wave-packet $(\mathbf{E}(t), \mathbf{H}(t))$ does not exceed 1 as $t \to \infty$, i.e.,

$$\int_{\{|x| \geq bt\}} |(\mathbf{E}(t), \mathbf{H}(t))|^2 dx \stackrel{t\to\infty}{\longrightarrow} 0 \text{ for all } b > 1. \tag{4.4}$$

Next it is shown that the potential energy and the energy of the curl-free part of the electromagnetic field decay in time mean, i.e.,

$$t^{-1} \int_0^t \left(\int_G \gamma^{-1} V(x, \mathbf{P}(s)) dx + \|(1-Q)(\mathbf{E}(s), \mathbf{H}(s))\|_X^2 \right) ds \stackrel{t\to\infty}{\longrightarrow} 0. \tag{4.5}$$

Here assumption (1.6), condition (1.8) and an $L^2 - L^6$-estimate for a vector potential are used. Theorem 3.2 and (4.5) yield the local decay of the electromagnetic field at least in time mean, i.e.,

$$t^{-1} \int_0^t \|(\mathbf{E}(s), \mathbf{H}(s))\|_{L^2(\Omega \cap B_R)}^2 ds \stackrel{t\to\infty}{\longrightarrow} 0 \text{ for all } R > 0. \tag{4.6}$$

The main step of the proof of (1.7) is a description of the asymptotic energy $\mathcal{E}_\infty$ in (2.23). Due to the fact that Ω is an exterior domain one has $\mathcal{E}_\infty > 0$ in general, even if condition (1.8) is satisfied. It is shown that for all $b > 1$

$$t^{-1} \int_{\{|x| \le bt\}} [SQ_0\chi_0(\mathbf{E}(t), \mathbf{H}(t))] \cdot Q_0\chi_0(\mathbf{E}(t), \mathbf{H}(t))dx \overset{t\to\infty}{\longrightarrow} \mathcal{E}_\infty, \tag{4.7}$$

where $S\mathbf{u} \overset{\text{def}}{=} (-x \wedge \underline{\mathbf{u}}_2, x \wedge \underline{\mathbf{u}}_1)$, $\chi_0 \in C^\infty(\mathbb{R}^3)$ is a cut-off function with supp $\chi_0 \subset \Omega$ and $\chi_0(x) = 1$ outside some bounded set. Furthermore, Q_0 denotes the orthogonal projector on the space of all $\mathbf{u} \in L^2(\mathbb{R}^3)$ with $\text{div } \underline{\mathbf{u}}_j = 0$. The proof of (4.7) relies on (4.4), (4.5), (4.6), and some weighted L^p-estimates for Q_0.

For this purpose the following additional assumptions are imposed on V, the set G and γ in what follows:

$$V(x, \mathbf{y}) \le K_2\mathbf{y}(\nabla_y V)(x, \mathbf{y}) \text{ for all } x \in G, \mathbf{y} \in \mathbb{R}^3 \tag{4.8}$$

with some $K_2 \in (0, \infty)$ independent of $x, \mathbf{y}$, and

$$\gamma \in L^{3/2}(G) \text{ and } (1 + |x|)\gamma \in L^{r_0}(G) \text{ with some } r_0 \in (3/2, \infty). \tag{4.9}$$

Finally it is assumed that the external current $\mathbf{j}$ is located in a fixed finite ball, i.e., there is some $R_1 > 0$ with

$$\mathbf{j}(t, x) = 0 \text{ for all } t \in (0, \infty), x \in \mathbb{R}^3 \setminus B_{R_1}. \tag{4.10}$$

First it is shown that the convergence in Theorem 3.2 is strong on bounded subsets of Ω.

LEMMA 4.1. *For all $R > 0$ one has*

$$\|Q(\mathbf{E}(t), \mathbf{H}(t))\|_{L^2(\Omega \cap B_R)} \overset{t\to\infty}{\longrightarrow} 0.$$

Proof. Each $\mathbf{u} \in (\ker B)^\perp$ satisfies

$$\text{div } (\underline{\mathbf{u}}_1) = 0, \quad \text{div } (\underline{\mathbf{u}}_2) = 0 \tag{4.11}$$

$$\text{with } \vec{n}\underline{\mathbf{u}}_1 = 0 \text{ on } \Gamma_2 \text{ and } \vec{n}\underline{\mathbf{u}}_2 = 0 \text{ on } \Gamma_1$$

in the sense that

$$\int_\Omega (\underline{\mathbf{u}}_1 \nabla\varphi + \underline{\mathbf{u}}_2 \nabla\psi)\, dx = 0 \text{ for all } \varphi \in C_0^\infty(\mathbb{R}^3 \setminus \overline{\Gamma_1}) \text{ and } \psi \in C_0^\infty(\mathbb{R}^3 \setminus \overline{\Gamma_2}).$$

This follows from the fact that $(\nabla\varphi, \nabla\psi) \in \ker B$ for all $\varphi \in C_0^\infty(\mathbb{R}^3 \setminus \overline{\Gamma_1})$ and $\psi \in C_0^\infty(\mathbb{R}^3 \setminus \overline{\Gamma_2})$.

Suppose $\mathbf{u} \in (\ker B)^\perp \cap D(B)$. Then $\underline{\mathbf{u}}_1 \in W_E$, whereas $\underline{\mathbf{u}}_2 \in W_H$. Therefore (4.11) and the compactness theorem in [7], a generalization of the result in [16] (see also [10] and [14]) implies that

$$(\ker B)^\perp \cap D(B) \text{ is compactly embedded in } L^2(\Omega \cap B_R) \text{ for all } R > 0. \tag{4.12}$$

Now, the result follows from Lemma 2.2(ii), Theorem 3.2, and (4.12). ∎

THEOREM 4.2. *Suppose $b > 1$. Then*

$$\int_{\{|x|\geq bt\}} |(\mathbf{E}(t), \mathbf{H}(t))|^2 dx \overset{t\to\infty}{\longrightarrow} 0.$$

Proof. The proof is based on an energy estimate. Let $g \in C^\infty(\mathbb{R})$ with $g(u) = 1$ for $u \geq (1+b)/2$ and $g(u) = 0$ for $u \leq 1$. For $R > R_1$ define

$$\mathcal{E}^{(R)}(t) \overset{\text{def}}{=} \int_\Omega g((t+R)^{-1}|x|) \left[|\mathbf{E}(t)|^2 + |\mathbf{H}(t)|^2\right] dx$$

$$+ \int_G \gamma^{-1} g((t+R)^{-1}|x|) \left(\alpha|\partial_t \mathbf{P}|^2 + 2V(x, \mathbf{P})\right) dx.$$

Then one obtains from the basic equations using (2.15) and assumption (4.10) for all $t \geq 0$

$$\frac{d}{dt}\mathcal{E}^{(R)}(t) = 2\left\langle g((t+R)^{-1}|x|)(\mathbf{E}(t), \mathbf{H}(t)), B(\mathbf{E}(t), \mathbf{H}(t)) - \mathcal{R}\partial_t \mathbf{P}(t) - (\mathbf{j}(t), 0)\right\rangle_X$$

$$-(t+R)^{-2} \int_\Omega |x| g'((t+R)^{-1}|x|) \left[|\mathbf{E}(t)|^2 + |\mathbf{H}(t)|^2\right] dx$$

$$+ \int_G 2g((t+R)^{-1}|x|) \left(\mathbf{E}\partial_t \mathbf{P} - \gamma^{-1}|\partial_t \mathbf{P}|^2\right) dx$$

$$-(t+R)^{-2} \int_G \gamma^{-1}|x| g'((t+R)^{-1}|x|) \left(\alpha|\partial_t \mathbf{P}|^2 + 2V(x, \mathbf{P})\right) dx$$

$$\leq 2\left\langle g((t+R)^{-1}|x|)(\mathbf{E}(t), \mathbf{H}(t)), B(\mathbf{E}(t), \mathbf{H}(t))\right\rangle_X$$

$$-(t+R)^{-2} \int_\Omega |x| g'((t+R)^{-1}|x|) \left[|\mathbf{E}(t)|^2 + |\mathbf{H}(t)|^2\right] dx$$

$$\leq 2(t+R)^{-1} \int_\Omega |x|^{-1} g'((t+R)^{-1}|x|) \mathbf{E}(t) \cdot (x \wedge \mathbf{H}(t)) dx$$

$$-(t+R)^{-2} \int_\Omega |x| g'((t+R)^{-1}|x|) \left[|\mathbf{E}(t)|^2 + |\mathbf{H}(t)|^2\right] dx.$$

Since $g(u) = 0$ for $u \leq 1$ and $g'(u) \geq 0$, it follows that $\frac{d}{dt}\mathcal{E}^{(R)}(t) \leq 0$ and hence

$$\mathcal{E}^{(R)}(t) \leq \mathcal{E}^{(R)}(0) \text{ for all } R > R_1. \tag{4.13}$$

Since $b > 1$, one has by (4.13) for all $R > R_1$

$$\limsup_{t\to\infty} \int_{\{|x|\geq bt\}} |(\mathbf{E}(t), \mathbf{H}(t))|^2 dx \leq \limsup_{t\to\infty} \int_{\{|x|\geq (b+1)(t+R)/2\}} |(\mathbf{E}(t), \mathbf{H}(t))|^2 dx$$

$$\leq \limsup_{t\to\infty} \mathcal{E}^{(R)}(t) \leq \mathcal{E}^{(R)}(0).$$

Since $g(0) = 0$, it follows that $\mathcal{E}^{(R)}(0) \overset{R\to\infty}{\longrightarrow} 0$. Hence the assertion follows from the previous estimate letting $R \to \infty$. □

In what follows let $R_0 > 0$ such that $\mathbb{R}^3 \setminus \Omega \subset B_{R_0} \overset{\text{def}}{=} \{x \in \mathbb{R}^3 : |x| < R_0\}$ and choose $\chi_0 \in C^\infty(\mathbb{R}^3)$ with

$$\text{supp } \chi_0 \subset \Omega \quad \text{and } \chi_0(x) = 1 \text{ on } \mathbb{R}^3 \setminus B_{R_0}. \tag{4.14}$$

For $\mathbf{w} \in X$ or $\mathbf{w} \in L^2(\mathbb{R}^3)$ define

$$\mathcal{C}_0\mathbf{w} \overset{\text{def}}{=} \left((\nabla\chi_0) \wedge \underline{\mathbf{w}}_2, -(\nabla\chi_0) \wedge \underline{\mathbf{w}}_1\right). \tag{4.15}$$

For convenience $\chi_0\mathbf{w}$ and $\mathcal{C}_0\mathbf{w}$ will be regarded as elements of $L^2(\mathbb{R}^3)$ by extending them by zero outside supp χ_0 if $\mathbf{w} \in X$.

In what follows $W_{E,0}$ denotes the space of all $\mathbf{F} \in W_E$ with curl $\mathbf{F} = 0$. Since $\nabla\varphi \in W_{E,0}$ for all $\varphi \in C_0^\infty(\Omega)$, one has

$$\text{div } \mathbf{A} = 0 \text{ for all } \mathbf{A} \in W_{E,0}^\perp. \tag{4.16}$$

By the boundedness of supp $\nabla\chi_0$ it follows from (4.16) that curl $(\chi_0\mathbf{A}) = (\nabla\chi_0) \wedge \mathbf{A} + \chi_0$ curl $\mathbf{A} \in L^2(\mathbb{R}^3)$ and div $(\chi_0\mathbf{A}) = (\nabla\chi_0)\cdot\mathbf{A} \in L^2(\mathbb{R}^3)$ for all $\mathbf{A} \in W_{E,0}^\perp \cap W_E$. Here $\chi_0\mathbf{A}$ is extended by zero on $\mathbb{R}^3\setminus$ supp χ_0 and χ_0 as in (4.14). From Sobolev's inequality [1] one obtains $\chi_0\mathbf{A} \in L^6(\mathbb{R}^3)$ and hence $\mathbf{A} \in L^6(\mathbb{R}^3 \setminus B_{R_0})$.

The aim of the following considerations is to prove the following estimate.

LEMMA 4.3. *There exists a constant $K_3 \in (0,\infty)$, such that for all $\mathbf{A} \in W_E \cap W_{E,0}^\perp$ the estimate*

$$\|\mathbf{A}\|_{L^2(\Omega\cap B_{R_0})} + \|\mathbf{A}\|_{L^6(\mathbb{R}^3\setminus B_{R_0})} \leq K_3 \|\ \textit{curl}\ \mathbf{A}\|_{L^2(\Omega)}$$

holds.

LEMMA 4.4. (i) *The set of all $\mathbf{F} \in W_{E,0}$ with bounded support is dense in $W_{E,0}$ (with respect to the $L^2(\Omega)$-norm).*

(ii) *Let $\mathbf{A} \in L^2(\Omega \cap B_{R_0}) \cap L^6(\mathbb{R}^3 \setminus B_{R_0})$ with*

$$\int_\Omega \mathbf{A}\ \textit{curl}\ \mathbf{h} dx = 0 \textit{ for all } \mathbf{h} \in C_0^\infty(\mathbb{R}^3 \setminus \overline{\Gamma_2}, \mathbb{C}^3) \tag{4.17}$$

and

$$\int_\Omega \mathbf{A}\mathbf{F} dx = 0 \textit{ for all } \mathbf{F} \in W_{E,0} \textit{ with bounded support.} \tag{4.18}$$

Then $\mathbf{A} = 0$.

Proof. (i) Suppose $\mathbf{F} \in W_{E,0}$. Since curl $\mathbf{F} = 0$, there exists some $\varphi \in L^6(\mathbb{R}^3 \setminus B_{R_0})$ with

$$\mathbf{F} = \nabla\varphi \text{ on } \mathbb{R}^3 \setminus B_{R_0}. \tag{4.19}$$

Let $\psi_1 \in C_0^\infty(B_2)$ with $\psi_1 = 1$ on B_1 and $\psi_n \overset{\text{def}}{=} \psi_1(x/n)$. Now define $\mathbf{F}_n(x) \overset{\text{def}}{=} \psi_n(x)\mathbf{F}(x)$ if $x \in \Omega \cap B_n$ and $\mathbf{F}_n(x) \overset{\text{def}}{=} \psi_n(x)\mathbf{F}(x) + \varphi(x)\nabla\psi_n(x)$ if $|x| \geq n$. Then $\mathbf{F}_n$ has bounded support and curl $\mathbf{F}_n = (\nabla\psi_n) \wedge \mathbf{F} + (\nabla\varphi) \wedge \nabla\psi_n = 0$ by (4.19). Since

also $\mathbf{F}_n = \mathbf{F}$ near $\partial\Omega$ it follows easily that $\mathbf{F}_n \in W_{E,0}$. Next it follows from Hölder's inequality that

$$\|\mathbf{F}_n - \mathbf{F}\|_{L^2(\Omega)} \leq \|(1-\psi_n)\mathbf{F}\|_{L^2(\Omega)} + \|\varphi\nabla\psi_n\|_{L^2(\Omega)}$$

$$\leq \|(1-\psi_n)\mathbf{F}\|_{L^2(\Omega)} + \|\varphi\|_{L^6(\{|x|>n\})}\|\nabla\psi_n\|_{L^3(\mathbb{R}^3)}$$

$$\leq \|(1-\psi_n)\mathbf{F}\|_{L^2(\Omega)} + \|\varphi\|_{L^6(\{|x|>n\})}\|\nabla\psi_n\|_{L^\infty(\mathbb{R}^3)}|B_{2n}|^{1/3}$$

$$\leq \|(1-\psi_n)\mathbf{F}\|_{L^2(\Omega)} + C_1\|\varphi\|_{L^6(\{|x|>n\})} \overset{n\to\infty}{\longrightarrow} 0$$

with some C_1 independent of n. This completes the proof of (i).

Next let $\mathbf{A} \in L^2(\Omega\cap B_{R_0}) \cap L^6(\mathbb{R}^3 \setminus B_{R_0})$ satisfy (4.17) and (4.18). Then one has in analogy to (4.16)

$$\text{curl } \mathbf{A} = 0 \text{ and } \text{ div } \mathbf{A} = 0 \text{ on } \Omega. \tag{4.20}$$

Since supp $\nabla\chi_0$ is bounded, it follows from (4.20) that curl $(\chi_0\mathbf{A}) = (\nabla\chi_0) \wedge \mathbf{A} \in L^{6/5}(\mathbb{R}^3)$ and div $(\chi_0\mathbf{A}) = (\nabla\chi_0)\cdot\mathbf{A} \in L^{6/5}(\mathbb{R}^3)$, where $\chi_0\mathbf{A}$ is extended by zero on $\mathbb{R}^3\setminus$ supp χ_0. From Lemma 1 in [8] one obtains $\chi_0\mathbf{A} \in L^2(\mathbb{R}^3)$ and hence $\mathbf{A} \in L^2(\Omega)$. By the definition of W_E and (4.17) we have $\mathbf{A} \in W_{E,0}$. Since $\mathbf{A} \in L^2(\Omega)$, (4.18) holds for all $\mathbf{F} \in W_{E,0}$ by assertion (i). But this means $\mathbf{A} \in W_{E,0}^{\perp}$ and therefore $\mathbf{A} = 0$. □

LEMMA 4.5. *Let $\{\mathbf{A}_n\}_{n\in\mathbb{N}}$ be a sequence in $W_E \cap W_{E,0}^{\perp}$ which is bounded in $L^2(\Omega\cap B_{R_0}) \cap L^6(\mathbb{R}^3 \setminus B_{R_0})$ such that $\{curl\ \mathbf{A}_n\}_{n\in\mathbb{N}}$ is precompact in $L^2(\Omega)$. Then $\{\mathbf{A}_n\}_{n\in\mathbb{N}}$ is precompact in $L^2(\Omega\cap B_{R_0}) \cap L^6(\mathbb{R}^3 \setminus B_{R_0})$.*

Proof. Let $\tilde{\Omega} \overset{\text{def}}{=} B_{2R_0} \cap \Omega$ and choose $\chi_1 \in C_0^\infty(B_{2R_0})$ with $\chi_1(x) = 1$ on B_{R_0}, in particular

$$\chi_1(x) = 1 \text{ on supp } (\nabla\chi_0). \tag{4.21}$$

Let $S_1 \overset{\text{def}}{=} \Gamma_1 \cup \partial B_{2R_0}$ and $S_2 \overset{\text{def}}{=} \Gamma_2 = \partial\tilde{\Omega} \setminus S_1$. Recall that $\mathbb{R}^3 \setminus \Omega \subset B_{R_0} \subset B_{2R_0}$. In analogy to the definition of W_E let $\mathcal{W}_E$ be the space of all $\mathbf{e} \in H_{curl}(\tilde{\Omega})$ with $\vec{n}\wedge\mathbf{e} = 0$ on S_1 in the sense that

$$\int_{\tilde{\Omega}} \mathbf{e} \text{ curl } \mathbf{f} - \mathbf{f} \text{ curl } \mathbf{e} dx = 0 \text{ for all } \mathbf{f} \in C_0^\infty(\mathbb{R}^3 \setminus \overline{S_2}, \mathbb{C}^3).$$

Now, it follows from the assumptions that

$$\{\chi_1\mathbf{A}_n\}_{n\in\mathbb{N}} \text{ is bounded in } \mathcal{W}_E. \tag{4.22}$$

Since $\mathbf{A}_n \in W_{E,0}^{\perp}$, one has also

$$\{\text{div } [\chi_1\mathbf{A}_n]\}_{n\in\mathbb{N}} = \{\mathbf{A}_n\nabla\chi_1\}_{n\in\mathbb{N}} \text{ is bounded in } L^2(\tilde{\Omega}) \tag{4.23}$$

and $\chi_1\vec{n}\mathbf{A} = 0$ on S_2, in the sense that

$$-\int_{\tilde{\Omega}} \chi_1\mathbf{A}_n\nabla\varphi dx = \int_{\tilde{\Omega}} (\text{div } [\chi_1\mathbf{A}_n])\varphi dx = \int_{\tilde{\Omega}} (\mathbf{A}_n\nabla\chi_1)\varphi dx$$

for all $\varphi \in C_0^\infty(\mathbb{R}^3 \setminus \overline{S_1})$.

Since the Lipschitz domain $\tilde{\Omega} = B_{2R_0} \cap \Omega$ and the decomposition of its boundary $\partial\tilde{\Omega} = S_1 \cup S_2$ satisfy the assumptions in [7], it follows from (4.22), (4.23), and the result in [7] that the sequence

$$\{\chi_1 \mathbf{A}_n\}_{n\in\mathbb{N}} \text{ is precompact in } L^2(\tilde{\Omega}) = L^2(B_{2R_0} \cap \Omega). \tag{4.24}$$

Let $\mathbf{f}_n(x) \overset{\text{def}}{=} \chi_0 \mathbf{A}_n(x)$ if $x \in \Omega$ and $\mathbf{f}_n(x) \overset{\text{def}}{=} 0$ if $x \in \mathbb{R}^3 \setminus \Omega$.

Next, (4.21), (4.24), and the compactness assumption on $\{\text{curl } \mathbf{A}_n\}_{n\in\mathbb{N}}$ imply that the sequences

$$\{\text{curl } \mathbf{f}_n\}_{n\in\mathbb{N}} = \{(\nabla\chi_0) \wedge \mathbf{A}_n + \chi_0 \text{ curl } \mathbf{A}_n\}_{n\in\mathbb{N}} \tag{4.25}$$

and

$$\{\text{div } \mathbf{f}_n\}_{n\in\mathbb{N}} = \{\mathbf{A}_n \nabla\chi_0\}_{n\in\mathbb{N}} \text{ are precompact in } L^2(\mathbb{R}^3). \tag{4.26}$$

Recall that supp $\chi_0 \subset \Omega$. By (4.25) and (4.26) it follows from Sobolev's inequality or directly from Lemma 1 in [8] that the sequence $(\mathbf{f}_n)_{n\in\mathbb{N}}$ is precompact in $L^6(\mathbb{R}^3)$ and hence

$$(\mathbf{A}_n)_{n\in\mathbb{N}} \text{ is precompact in } L^6(\Omega \setminus B_{R_0}) \tag{4.27}$$

since $\chi_0(x) = 1$ for $|x| > R_0$. □

Proof of Lemma 4.3. Suppose that the estimate was not correct, i.e., there exists a sequence $\mathbf{A}_n \in W_E \cap W_{E,0}^\perp, n \in \mathbb{N}$, with

$$1 = \|\mathbf{A}_n\|_{L^2(\Omega\cap B_{R_0})} + \|\mathbf{A}_n\|_{L^6(\mathbb{R}^3\setminus B_{R_0})} \geq n\| \text{ curl } \mathbf{A}_n\|_{L^2} \text{ for all } n \in \mathbb{N}. \tag{4.28}$$

By Lemma 4.5 the sequence $\{\mathbf{A}_n\}_{n\in\mathbb{N}}$ is precompact in $L^2(\Omega \cap B_{R_0}) \cap L^6(\mathbb{R}^3 \setminus B_{R_0})$, i.e., there exist $\mathbf{A} \in L^2(\Omega \cap B_{R_0}) \cap L^6(\mathbb{R}^3 \setminus B_{R_0})$ and a subsequence $\mathbf{A}_{n_k}, k \in \mathbb{N}$, with

$$\|\mathbf{A}_{n_k} - \mathbf{A}\|_{L^2(\Omega\cap B_{R_0})} + \|\mathbf{A}_{n_k} - \mathbf{A}\|_{L^6(\mathbb{R}^3\setminus B_{R_0})} \overset{k\to\infty}{\longrightarrow} 0, \tag{4.29}$$

in particular

$$\|\mathbf{A}\|_{L^2(\Omega\cap B_{R_0})} + \|\mathbf{A}\|_{L^6(\mathbb{R}^3\setminus B_{R_0})} = 1. \tag{4.30}$$

From (4.28) and (4.29) it follows that

$$\int_\Omega \mathbf{A} \text{ curl } \mathbf{h} dx = \lim_{k\to\infty} \int_\Omega \mathbf{A}_{n_k} \text{ curl } \mathbf{h} dx \tag{4.31}$$

$$= \lim_{k\to\infty} \int_\Omega \mathbf{h} \text{ curl } \mathbf{A}_{n_k} dx = 0 \text{ for all } \mathbf{h} \in C_0^\infty(\mathbb{R}^3 \setminus \overline{\Gamma_2}, \mathbb{C}^3).$$

Furthermore

$$\int_\Omega \mathbf{A}\mathbf{F} dx = \lim_{k\to\infty} \int_\Omega \mathbf{A}_{n_k} \mathbf{F} dx = 0 \tag{4.32}$$

for all $\mathbf{F} \in W_{E,0}$ with bounded support. Now (4.31), (4.32), and Lemma 4.4(ii) would imply $\mathbf{A} = 0$. This contradicts (4.30). □

The aim of the following considerations is to show decay of the potential energy and the local electromagnetic energy at least in time mean, i.e., for all $R > 0$

$$t^{-1} \int_0^t \left(\int_G \gamma^{-1} V(x, \mathbf{P}(s)) dx + \|(\mathbf{E}(s), \mathbf{H}(s))\|^2_{L^2(\Omega \cap B_R)} \right) ds \stackrel{t \to \infty}{\longrightarrow} 0.$$

LEMMA 4.6. *There holds*

$$t^{-1} \int_0^t \langle Q(\mathbf{E}(s), \mathbf{H}(s)), \mathcal{R}\mathbf{P}(s) \rangle_X ds \stackrel{t \to \infty}{\longrightarrow} 0.$$

Proof. Let $\mathbf{u}(t) \stackrel{\text{def}}{=} Q(\mathbf{E}(t), \mathbf{H}(t))$ and $\mathbf{A}(t) \stackrel{\text{def}}{=} \int_0^t \underline{\mathbf{u}}_1(s) ds$. Since $\mathbf{u}(t) \in (\ker B)^\perp$ one has $\mathbf{A}(t) \in W^\perp_{E,0}$. With

$$\text{curl } \underline{\mathbf{u}}_1(s) = -\underline{(B\mathbf{u}(s))}_2 = -\underline{[B(\mathbf{E}(t), \mathbf{H}(t))]}_2 = \text{ curl } \mathbf{E}(s) = -\partial_t \mathbf{H}(s)$$

one gets by using Lemma 4.3

$$\|\mathbf{A}(t)\|_{L^2(\Omega \cap B_{R_0})} + \|\mathbf{A}(t)\|_{L^6(\mathbb{R}^3 \setminus B_{R_0})} \leq K_3 \| \text{ curl } \mathbf{A}(t)\|_{L^2(\Omega)}$$

$$= K_3 \| \int_0^t \text{ curl } \underline{\mathbf{u}}_1(s) ds \|_{L^2(\Omega)} = K_3 \|\mathbf{H}(0) - \mathbf{H}(t)\|_{L^2(\Omega)}.$$

Now, it follows from Lemma 2.2 and the previous estimate that

$$\|\mathbf{A}(t)\|_{L^2(\Omega \cap B_{R_0})} + \|\mathbf{A}(t)\|_{L^6(\mathbb{R}^3 \setminus B_{R_0})} \leq C_1 \text{ for all } t \in (0, \infty) \tag{4.33}$$

with some constant C_1 independent of t. Next,

$$t^{-1} \int_0^t \langle Q(\mathbf{E}(s), \mathbf{H}(s)), \mathcal{R}\mathbf{P}(s) \rangle_X ds = t^{-1} \int_0^t \int_G \underline{\mathbf{u}}_1(s) \mathbf{P}(s) dx ds \tag{4.34}$$

$$= t^{-1} \int_0^t \int_G \partial_t \mathbf{A}(s) \mathbf{P}(s) dx ds = t^{-1} \int_G \mathbf{A}(t) \mathbf{P}(t) dx - t^{-1} \int_0^t \int_G \mathbf{A}(s) \partial_t \mathbf{P}(s) dx ds$$

$$\leq C_1 t^{-1} \left(\|\mathbf{P}(t)\|_{L^2(G)} + \|\mathbf{P}(t)\|_{L^{6/5}(G)} \right)$$

$$+ C_1 t^{-1} \int_0^t \left(\|\partial_t \mathbf{P}(s)\|_{L^2(G)} + \|\partial_t \mathbf{P}(s)\|_{L^{6/5}(G)} \right) ds$$

$$\leq C_1 t^{-1} \left(\|\gamma^{1/2}\|_{L^\infty(G)} + \|\gamma^{1/2}\|_{L^3(G)} \right) \|\mathbf{P}(t)\|_{\mathcal{G}}$$

$$+ C_1 t^{-1} \int_0^t \left(\|\gamma^{1/2}\|_{L^\infty(G)} + \|\gamma^{1/2}\|_{L^3(G)} \right) \|\partial_t \mathbf{P}(s)\|_{\mathcal{G}} ds$$

by assumption (4.9) and Hölder's inequality. Next Lemma 2.2 yields

$$\|\mathbf{P}(t)\|_{\mathcal{G}} \leq \|\mathbf{P}_0\|_{\mathcal{G}} + \int_0^t \|\partial_s \mathbf{P}(s)\|_{\mathcal{G}} ds$$

$$\leq C_2 + C_2 t^{1/2} \|\partial_s \mathbf{P}\|_{L^2((0,\infty),\mathcal{G})} \leq C_3 + C_3 t^{1/2}.$$

With (4.34) and Lemma 2.2 again one obtains

$$t^{-1} \int_0^t \langle Q(\mathbf{E}(s), \mathbf{H}(s)), \mathcal{R}\mathbf{P}(s) \rangle_X ds$$

$$\leq C_4(t^{-1} + t^{-1/2}) + C_4 t^{-1/2} \|\partial_s \mathbf{P}\|_{L^2((0,\infty),\mathcal{G})} \stackrel{t\to\infty}{\longrightarrow} 0. \qquad \square$$

LEMMA 4.7. *There holds*

$$t^{-1} \int_0^t \left(\int_G \gamma^{-1} V(x, \mathbf{P}(s)) dx + \|(1-Q)(\mathbf{E}(t), \mathbf{H}(t))\|_X^2 \right) ds \stackrel{t\to\infty}{\longrightarrow} 0,$$

in particular

$$\lim_{t\to\infty} t^{-1} \int_0^t \|(\mathbf{E}(s), \mathbf{H}(s))\|_X^2 ds = \lim_{t\to\infty} t^{-1} \int_0^t \|Q(\mathbf{E}(s), \mathbf{H}(s))\|_X^2 ds = \mathcal{E}_\infty,$$

where $\mathcal{E}_\infty$ *as in* (2.23).

Proof. It follows from Lemma 2.2 and (4.3) that

$$\|(1-Q)(\mathbf{E}(t), \mathbf{H}(t))\|_X^2 = \langle (\mathbf{E}(t), \mathbf{H}(t)), (1-Q)[(\mathbf{J}(t), 0) - \mathcal{R}\mathbf{P}(t)] \rangle_X$$

$$= \langle Q(\mathbf{E}(t), \mathbf{H}(t)), \mathcal{R}\mathbf{P}(t) \rangle_X + \langle (\mathbf{E}(t), \mathbf{H}(t)), (1-Q)(\mathbf{J}(t), 0) \rangle_X - \int_G \mathbf{E}(t)\mathbf{P}(t) dx$$

$$\leq \langle Q(\mathbf{E}(t), \mathbf{H}(t)), \mathcal{R}\mathbf{P}(t) \rangle_X + C_1 \|\mathbf{J}(t)\|_{L^2(\Omega)}$$

$$- \int_G \gamma^{-1} \left[\alpha \partial_t^2 \mathbf{P}(t) + \partial_t \mathbf{P}(t) + (\nabla_y V)(x, \mathbf{P}(t)) \right] \mathbf{P} dx$$

$$\leq \langle Q(\mathbf{E}(t), \mathbf{H}(t)), \mathcal{R}\mathbf{P}(t) \rangle_X + C_1 \|\mathbf{J}(t)\|_{L^2(\Omega)}$$

$$- \int_G \gamma^{-1} \left(\alpha \partial_t^2 \mathbf{P}(t) + \partial_t \mathbf{P}(t) \right) \mathbf{P}(t) dx - K_2^{-1} \int_G \gamma^{-1} V(x, \mathbf{P}(t)) dx$$

by assumption (4.8). Now,

$$t^{-1} \int_0^t \left(\|(1-Q)(\mathbf{E}(s), \mathbf{H}(s))\|_X^2 + K_2^{-1} \int_G \gamma^{-1} V(x, \mathbf{P}(s)) dx \right) ds \tag{4.35}$$

$$\leq t^{-1} \int_0^t \left(\langle Q(\mathbf{E}(s), \mathbf{H}(s)), \mathcal{R}\mathbf{P}(s) \rangle_X + C_1 \|\mathbf{J}(s)\|_{L^2(\Omega)} \right) ds$$

$$+ C_2/t - 1/2 t^{-1} \int_G \gamma^{-1} |\mathbf{P}(t)|^2 dx - t^{-1} \int_G \alpha \gamma^{-1} \partial_t \mathbf{P}(t) \mathbf{P}(t) dx$$

$$+t^{-1}\int_0^t\int_G \alpha\gamma^{-1}|\partial_t\mathbf{P}(s)|^2dxds$$

$$\leq t^{-1}\int_0^t \left(\langle Q(\mathbf{E}(s),\mathbf{H}(s)),\mathcal{R}\mathbf{P}(s)\rangle_X + C_1\|\mathbf{J}(s)\|_{L^2(\Omega)}\right)ds$$

$$+C_3/t + C_3t^{-1}\|\partial_t\mathbf{P}(t)\|^2_{\mathcal{G}} + t^{-1}\|\alpha^{1/2}\partial_t\mathbf{P}\|^2_{L^2((0,\infty),\mathcal{G})}$$

$$\leq t^{-1}\int_0^t \left(\langle Q(\mathbf{E}(s),\mathbf{H}(s)),\mathcal{R}\mathbf{P}(s)\rangle_X + C_1\|\mathbf{J}(s)\|_{L^2(\Omega)}\right)ds + C_4/t$$

by Lemma 2.2(i) again.

In the previous estimates C_j are constants independent of t. Now, it follows from Lemma 4.6 and (4.35) that

$$t^{-1}\int_0^t \left(\int_G \gamma^{-1}V(x,\mathbf{P}(s))dx + \|(1-Q)(\mathbf{E}(t),\mathbf{H}(t))\|^2_X\right)ds \stackrel{t\to\infty}{\longrightarrow} 0. \tag{4.36}$$

Since

$$t^{-1}\int_0^t \|\alpha^{1/2}\partial_t\mathbf{P}(s)\|^2_{\mathcal{G}}ds \leq t^{-1}\|\alpha^{1/2}\partial_t\mathbf{P}\|^2_{L^2((0,\infty),\mathcal{G})}dxds \stackrel{t\to\infty}{\longrightarrow} 0$$

by Lemma 2.2(i), (2.23) and (4.36) yield

$$\lim_{t\to\infty} t^{-1}\int_0^t \|Q(\mathbf{E}(s),\mathbf{H}(s))\|^2_X ds = \lim_{t\to\infty} t^{-1}\int_0^t \|(\mathbf{E}(s),\mathbf{H}(s))\|^2_X ds = \mathcal{E}_\infty,$$

which completes the proof. □

From Lemma 4.1 and the previous lemma one now easily obtains the following.

COROLLARY 4.8. *For all $R>0$ one has*

$$t^{-1}\int_0^t \|(\mathbf{E}(s),\mathbf{H}(s))\|_{L^2(\Omega\cap B_R)}ds \stackrel{t\to\infty}{\longrightarrow} 0.$$

In what follows let

$$D(B_0) \stackrel{\text{def}}{=} H_{curl}(\mathbb{R}^3)\times H_{curl}(\mathbb{R}^3) \text{ and } B_0(\mathbf{e},\mathbf{h}) \stackrel{\text{def}}{=} (\operatorname{curl}\mathbf{h}, -\operatorname{curl}\mathbf{e}).$$

Furthermore, let Q_0 be the orthogonal projector on $(\ker B_0)^\perp$, which consists of all $\mathbf{u}\in L^2(\mathbb{R}^3)$ with $\operatorname{div}\underline{\mathbf{u}}_j = 0$. The following estimate will be used in the proof of (4.7).

LEMMA 4.9. *Let $s\in[0,1]$. Then there exists a constant $K_1\in(0,\infty)$ such that*

$$|\langle(1+|x|)^s\mathbf{f}, Q_0\mathbf{g}\rangle_{L^2(\mathbb{R}^3)}| \leq K_1\|\mathbf{f}\|_{H^1(\mathbb{R}^3)}\|(1+|x|)^s\mathbf{g}\|_{L^{q_0}(\mathbb{R}^3)}$$

for all $\mathbf{f}\in H^1(\mathbb{R}^3)$ and $\mathbf{g}\in L^2(\mathbb{R}^3)$ with $(1+|x|)^s\mathbf{f}\in L^2(\mathbb{R}^3)$ and $(1+|x|)^s\mathbf{g}\in L^{q_0}(\mathbb{R}^3)$. Here $1/q_0 = 1/(2r_0)+1/2$, where r_0 is as in assumption (4.9).

Proof. By a standard density argument it suffices to consider $\mathbf{f}, \mathbf{g} \in C_0^\infty(\mathbb{R}^3)$. Recall that $2r_0 > 3$. Let $p_1 \stackrel{\text{def}}{=} (\frac{1-s}{3} - \frac{1-s}{2r_0})^{-1} \in (\frac{3}{1-s}, \infty]$ ($p_1 = \infty$ for $s = 1$) and $p_2 \stackrel{\text{def}}{=} (\frac{s}{3} - \frac{s}{2r_0})^{-1}$. Since $(s-1)p_1 < -3$ and $sp_2 > 3$ one has

$$(1+|x|)^{s-1} \in L^{p_1}(\mathbb{R}^3) \text{ and } (1+|x|)^{-s} \in L^{p_2}(\mathbb{R}^3). \tag{4.37}$$

Now $\mathbf{F} \stackrel{\text{def}}{=} (1+|x|)^s Q_0 \mathbf{f} - Q_0\left((1+|x|)^s \mathbf{f}\right)$ obeys

$$B_0\mathbf{F} = -s(1+|x|)^{s-1}|x|^{-1} S Q_0 \mathbf{f} + (1+|x|)^s B_0 \mathbf{f} - B_0\left((1+|x|)^s \mathbf{f}\right)$$

$$= -s(1+|x|)^{s-1}|x|^{-1} S[Q_0\mathbf{f} - \mathbf{f}],$$

where $S\mathbf{w} \stackrel{\text{def}}{=} (-x \wedge \underline{\mathbf{w}}_2, x \wedge \underline{\mathbf{w}}_1)$. Hence

$$\|(1+|x|)^{1-s} B_0\mathbf{F}\|_{L^2(\mathbb{R}^3)} \le s\|(Q_0 - 1)\mathbf{f}\|_{L^2(\mathbb{R}^3)} \le s\|\mathbf{f}\|_{L^2(\mathbb{R}^3)}.$$

A similar estimate using $\text{div } \underline{(Q_0\mathbf{f})}_j = 0$ yields

$$\|(1+|x|)^{1-s} \text{ div } \underline{\mathbf{F}}_j\|_{L^2(\mathbb{R}^3)} \le s\|\underline{(Q_0\mathbf{f})}_j\|_{L^2(\mathbb{R}^3)} \le s\|\mathbf{f}\|_{L^2(\mathbb{R}^3)}.$$

With $1/q \stackrel{\text{def}}{=} 1/p_1 + 1/2$ we obtain by (4.37) and Hölder's inequality

$$\| \text{ curl } \underline{\mathbf{F}}_j\|_{L^q(\mathbb{R}^3)} \le C_1\|B_0\mathbf{F}\|_{L^q(\mathbb{R}^3)}$$

$$\le C_1\|(1+|x|)^{s-1}\|_{L^{p_1}(\mathbb{R}^3)}\|(1+|x|)^{1-s} B_0\mathbf{F}\|_{L^2(\mathbb{R}^3)} \le C_2\|\mathbf{f}\|_{L^2(\mathbb{R}^3)}$$

and

$$\| \text{ div } \underline{\mathbf{F}}_j\|_{L^q(\mathbb{R}^3)} \le C_2\|\mathbf{f}\|_{L^2(\mathbb{R}^3)}.$$

By Sobolev's inequality [1] or directly by Lemma 1 in [8] one obtains

$$\|\mathbf{F}\|_{L^r(\mathbb{R}^3)} \le C_3\|D\mathbf{F}\|_{L^q(\mathbb{R}^3)} \tag{4.38}$$

$$\le C_4\left(\|\text{curl } \underline{\mathbf{F}}_j\|_{L^q(\mathbb{R}^3)} + \| \text{ div } \underline{\mathbf{F}}_j\|_{L^q(\mathbb{R}^3)}\right) \le C_3\|\mathbf{f}\|_{L^2(\mathbb{R}^3)}.$$

Here $r \stackrel{\text{def}}{=} (\frac{1}{q} - \frac{1}{3})^{-1} = (1/6 + \frac{1-s}{3} - \frac{1-s}{2r_0})^{-1} \in (2, 6)$. Now, $\frac{1}{r} + \frac{1}{p_2} + \frac{1}{q_0} = 1$. Therefore (4.37), (4.38), and Hölder's inequality yield

$$|\langle \mathbf{F}, \mathbf{g}\rangle_{L^2(\mathbb{R}^3)}| \le \|\mathbf{F}\|_{L^r(\mathbb{R}^3)}\|(1+|x|)^{-s}\|_{L^{p_2}(\mathbb{R}^3)}\|(1+|x|)^s \mathbf{g}\|_{L^{q_0}(\mathbb{R}^3)} \tag{4.39}$$

$$\le C_5\|\mathbf{f}\|_{L^2(\mathbb{R}^3)}\|(1+|x|)^s \mathbf{g}\|_{L^{q_0}(\mathbb{R}^3)}.$$

Since $q_0 \ge 6/5$ one has $q_0^* \le 6$. Therefore, it follows from Hölder's inequality, (4.39), and the embedding $H^1(\mathbb{R}^3) \hookrightarrow L^{q_0^*}(\mathbb{R}^3)$ that

$$|\langle (1+|x|)^s \mathbf{f}, Q_0\mathbf{g}\rangle_{L^2(\mathbb{R}^3)}| \le |\langle Q_0\mathbf{f}, (1+|x|)^s \mathbf{g}\rangle_{L^2(\mathbb{R}^3)}| + |\langle \mathbf{F}, \mathbf{g}\rangle_{L^2(\mathbb{R}^3)}| \tag{4.40}$$

$$\leq C_5\|Q_0\mathbf{f}\|_{H^1(\mathbb{R}^3)}\|(1+|x|)^s\mathbf{g}\|_{L^{q_0}(\mathbb{R}^3)} + |\langle\mathbf{F},\mathbf{g}\rangle_{L^2(\mathbb{R}^3)}|$$

$$\leq C_6\|\mathbf{f}\|_{H^1(\mathbb{R}^3)}\|(1+|x|)^s\mathbf{g}\|_{L^{q_0}(\mathbb{R}^3)} \qquad \square$$

Since supp $\chi_0 \subset \Omega$, Lemma 2.2 yields $\chi_0(\mathbf{E}(\cdot),\mathbf{H}(\cdot)) \in L^\infty((0,\infty), D(B_0))$ and hence

$$Q_0\chi_0(\mathbf{E}(\cdot),\mathbf{H}(\cdot)) \in L^\infty((0,\infty), D(B_0)\cap(\ker\, B_0)^\perp) \tag{4.41}$$

$$\subset L^\infty((0,\infty), H^1(\mathbb{R}^3)),$$

where $\chi_0 \in C^\infty(\mathbb{R}^3)$ as in (4.14).

LEMMA 4.10. *There holds* (i)

$$t^{-1}\int_0^t \|(1-Q_0)\chi_0(\mathbf{E}(s),\mathbf{H}(s))\|_{L^2(\mathbb{R}^3)}ds \overset{t\to\infty}{\longrightarrow} 0.$$

(ii)

$$t^{-1}\int_0^t \|Q_0\chi_0(\mathbf{E}(s),\mathbf{H}(s))\|^2_{L^2(\mathbb{R}^3)}ds \overset{t\to\infty}{\longrightarrow} \mathcal{E}_\infty, \textit{ and}$$

(iii)

$$t^{-1}\int_0^t \|Q_0\chi_0(\mathbf{E}(s),\mathbf{H}(s))\|_{L^2(\{|x|\geq as\})}ds \overset{t\to\infty}{\longrightarrow} 0 \textit{ for all } a\in(1,\infty).$$

Proof. First the following estimate is proved:

$$\|(1-Q_0)\chi_0\mathbf{u}\|_{L^2(\mathbb{R}^3)} \leq K_1\left(\|(1-Q)\mathbf{u}\|_X + \|\mathbf{u}\|_{L^2(\Omega\cap B_{R_0})}\right) \tag{4.42}$$

for all $\mathbf{u}\in X$ with some constant independent of $\mathbf{u}$. For this purpose let $\mathbf{u}\in X$ and define $\mathbf{f} \overset{\text{def}}{=} (1-Q_0)\chi_0\mathbf{u} \in \ker\, B_0$, i.e., curl $\underline{\mathbf{f}}_j = 0$ on $\mathbb{R}^3$. Hence there exist $\varphi_j \in L^6(\mathbb{R}^3)$ with $\nabla\varphi_j \in L^2(\mathbb{R}^3)$ such that

$$\underline{\mathbf{f}}_j = \nabla\varphi_j. \tag{4.43}$$

Hence

$$\|\mathbf{f}\|^2_{L^2(\mathbb{R}^3)} = \langle\chi_0\mathbf{u},\mathbf{f}\rangle_{L^2(\mathbb{R}^3)} = \int_{\mathbb{R}^3}\chi_0\cdot(\underline{\mathbf{u}}_1\nabla\varphi_1 + \underline{\mathbf{u}}_2\nabla\varphi_2)\,dx$$

$$= \langle\mathbf{u},(\nabla[\chi_0\varphi_1],\nabla[\chi_0\varphi_2])\rangle_X - \int_{\mathbb{R}^3}(\underline{\mathbf{u}}_1(\nabla\chi_0)\varphi_1 - \underline{\mathbf{u}}_2(\nabla\chi_0)\varphi_2)\,dx.$$

Since $(\nabla[\chi_0\varphi_1],\nabla[\chi_0\varphi_2]) \in \ker B$ and supp $(\nabla\chi_0)$ is bounded, it follows that

$$\|(1-Q_0)\chi_0\mathbf{u}\|^2_{L^2(\mathbb{R}^3)} = \|\mathbf{f}\|^2_{L^2(\mathbb{R}^3)} \tag{4.44}$$

$$\leq C_1\|(1-Q)\mathbf{u}\|_X\left(\|\nabla(\chi_0\varphi_1)\|_{L^2(\mathbb{R}^3)} + \|\nabla(\chi_0\varphi_2)\|_{L^2(\mathbb{R}^3)}\right)$$

$$+C_1\|\mathbf{u}\|_{L^2(\Omega\cap B_{R_0})}\left(\|\varphi_1\|_{L^2(B_{R_0})}+\|\varphi_2\|_{L^2(B_{R_0})}\right)$$

$$\leq C_1\left(\|(1-Q)\mathbf{u}\|_X+\|\mathbf{u}\|_{L^2(\Omega\cap B_{R_0})}\right)$$

$$\left(\|\nabla\varphi_1\|_{L^2(\mathbb{R}^3)}+\|\varphi_1\|_{L^6(\mathbb{R}^3)}+\|\nabla\varphi_2\|_{L^2(\mathbb{R}^3)}+\|\varphi_1\|_{L^6(\mathbb{R}^3)}\right)$$

$$\leq C_2\left(\|(1-Q)\mathbf{u}\|_X+\|\mathbf{u}\|_{L^2(\Omega\cap B_{R_0})}\right)\|\mathbf{f}\|_{L^2(\mathbb{R}^3)}.$$

This completes the proof of (4.42). Now, assertion (i) follows immediately from Lemma 4.7, Corollary 4.8, and inequality (4.42).

Next, part (i), Corollary 4.8, and Lemma 4.7 yield by the boundedness of supp $(1-\chi_0)$

$$\lim_{t\to\infty}t^{-1}\int_0^t\|Q_0\chi_0(\mathbf{E}(s),\mathbf{H}(s))\|^2_{L^2(\mathbb{R}^3)}ds=\lim_{t\to\infty}t^{-1}\int_0^t\|\chi_0(\mathbf{E}(s),\mathbf{H}(s))\|^2_{L^2(\mathbb{R}^3)}ds$$

$$=\lim_{t\to\infty}t^{-1}\int_0^t\|(\mathbf{E}(s),\mathbf{H}(s))\|^2_X ds=\mathcal{E}_\infty,$$

whence (ii). Finally, part (iii) follows from (i) and Theorem 4.2. □

Next (4.7) is proved.

THEOREM 4.11. *Suppose $g\in C_0^\infty(\mathbb{R})$ with $g(u)=1$ on a neighborhood of $[0,1]$. Then*

$$t^{-1}\langle Sg(|x|/t)Q_0\chi_0(\mathbf{E}(t),\mathbf{H}(t)),Q_0\chi_0(\mathbf{E}(t),\mathbf{H}(t))\rangle_{L^2(\mathbb{R}^3)}\overset{t\to\infty}{\longrightarrow}\mathcal{E}_\infty,$$

where $\mathcal{E}_\infty$ is as in (2.23) *and $S\mathbf{u}\overset{\text{def}}{=}(-x\wedge\underline{\mathbf{u}}_2,x\wedge\underline{\mathbf{u}}_1)$.*

Proof. Define

$$F(t)\overset{\text{def}}{=}\langle Sg(|x|/t)Q_0\chi_0(\mathbf{E}(t),\mathbf{H}(t)),Q_0\chi_0(\mathbf{E}(t),\mathbf{H}(t))\rangle_{L^2(\mathbb{R}^3)}.$$

Then

$$F'(t)=2\,\langle Sg(|x|/t)Q_0\chi_0(\mathbf{E}(t),\mathbf{H}(t)),\tag{4.45}$$

$$Q_0\chi_0\left(B(\mathbf{E}(t),\mathbf{H}(t))-(\mathbf{j}(t),0)-\mathcal{R}\partial_t\mathbf{P}(t)\right)\rangle_{L^2(\mathbb{R}^3)}$$

$$-t^{-2}\,\langle S|x|g'(|x|/t)Q_0\chi_0(\mathbf{E}(t),\mathbf{H}(t)),Q_0\chi_0(\mathbf{E}(t),\mathbf{H}(t))\rangle_{L^2(\mathbb{R}^3)}$$

$$=\sum_{j=0}^2 h_j(t)+2\,\langle Sg(|x|/t)Q_0\chi_0(\mathbf{E}(t),\mathbf{H}(t)),B_0Q_0\chi_0(\mathbf{E}(t),\mathbf{H}(t))\rangle_{L^2(\mathbb{R}^3)}$$

$$-t^{-2}\,\langle S|x|g'(|x|/t)Q_0\chi_0(\mathbf{E}(t),\mathbf{H}(t)),Q_0\chi_0(\mathbf{E}(t),\mathbf{H}(t))\rangle_{L^2(\mathbb{R}^3)}$$

by (2.15). Here

$$h_0(t) \stackrel{\text{def}}{=} -2\,\langle Sg(|x|/t)Q_0\chi_0(\mathbf{E}(t),\mathbf{H}(t)), Q_0\chi_0(\mathbf{j}(t),0)\rangle_{L^2(\mathbb{R}^3)}\,, \tag{4.46}$$

$$h_1(t) \stackrel{\text{def}}{=} -2\,\langle Sg(|x|/t)Q_0\chi_0(\mathbf{E}(t),\mathbf{H}(t)), Q_0\chi_0\mathcal{R}\partial_t\mathbf{P}(t)\rangle_{L^2(\mathbb{R}^3)}\,, \tag{4.47}$$

$$h_2(t) \stackrel{\text{def}}{=} -2\,\langle Sg(|x|/t)Q_0\chi_0(\mathbf{E}(t),\mathbf{H}(t)), Q_0\mathcal{C}_0(\mathbf{E}(t),\mathbf{H}(t))\rangle_{L^2(\mathbb{R}^3)}\,, \tag{4.48}$$

where $\mathcal{C}_0$ is as in (4.15).

For $\mathbf{u} \in (\ker B_0)^\perp \cap D(B_0) \subset H^1(\mathbb{R}^3)$ one has $\operatorname{div} \underline{\mathbf{u}}_j = 0$. Therefore, it follows from the identity $x \wedge \operatorname{curl} \mathbf{a} = \nabla(x\mathbf{a}) - \mathbf{a} - (x\nabla)\mathbf{a}$ that

$$\begin{aligned}
&\langle Sg(|x|/t)\mathbf{u}, B_0\mathbf{v}\rangle_{L^2(\mathbb{R}^3)} + \langle Sg(|x|/t)B_0\mathbf{u}, \mathbf{v}\rangle_{L^2(\mathbb{R}^3)} \\
&= \int_{\mathbb{R}^3} g(|x|/t)\mathbf{u}\cdot(x\wedge \operatorname{curl}\underline{\mathbf{v}}_1, x\wedge \operatorname{curl}\underline{\mathbf{v}}_2)dx \\
&\quad + \int_{\mathbb{R}^3} g(|x|/t)(x\wedge \operatorname{curl}\underline{\mathbf{u}}_1, x\wedge \operatorname{curl}\underline{\mathbf{u}}_2)\cdot\mathbf{v}dx \\
&= \int_{\mathbb{R}^3} g(|x|/t)\mathbf{u}\cdot(\nabla[x\underline{\mathbf{v}}_1], \nabla[x\underline{\mathbf{v}}_2])\,dx + \int_{\mathbb{R}^3} g(|x|/t)\,(\nabla[x\underline{\mathbf{u}}_1], \nabla[x\underline{\mathbf{u}}_2])\cdot\mathbf{v}dx \\
&\quad -2\int_{\mathbb{R}^3} g(|x|/t)\mathbf{u}\cdot\mathbf{v}dx - \int_{\mathbb{R}^3} g(|x|/t)(x\nabla)[\mathbf{u}\mathbf{v}]dx \\
&= -2t^{-1}\left\langle \tilde{S}g'(|x|/t)\mathbf{u}, \mathbf{v}\right\rangle_{L^2(\mathbb{R}^3)} + \left\langle [g(|x|/t) + t^{-1}|x|g'(|x|/t)]\mathbf{u}, \mathbf{v}\right\rangle_{L^2(\mathbb{R}^3)}
\end{aligned}$$

with $\tilde{S}\mathbf{u} \stackrel{\text{def}}{=} |x|^{-1}\,([x\underline{\mathbf{u}}_1]x, [x\underline{\mathbf{u}}_2]x)$.
Hence

$$\begin{aligned}
&2\,\langle Sg(|x|/t)Q_0\chi_0(\mathbf{E}(t),\mathbf{H}(t)), B_0Q_0\chi_0(\mathbf{E}(t),\mathbf{H}(t))\rangle_{L^2(\mathbb{R}^3)} \\
&= \left\langle [g(|x|/t) + t^{-1}|x|g'(|x|/t)]Q_0\chi_0(\mathbf{E}(t),\mathbf{H}(t)), Q_0\chi_0(\mathbf{E}(t),\mathbf{H}(t))\right\rangle_{L^2(\mathbb{R}^3)} \\
&\quad -2t^{-1}\left\langle \tilde{S}g'(|x|/t)Q_0\chi_0(\mathbf{E}(t),\mathbf{H}(t)), Q_0\chi_0(\mathbf{E}(t),\mathbf{H}(t))\right\rangle_{L^2(\mathbb{R}^3)}.
\end{aligned}$$

With (4.45)–(4.48) it follows that

$$F'(t) = \|Q_0\chi_0(\mathbf{E}(t),\mathbf{H}(t))\|^2_{L^2(\mathbb{R}^3)} + \sum_{j=0}^{3} h_j(t), \tag{4.49}$$

where

$$h_3(t) \tag{4.50}$$

$$\stackrel{\text{def}}{=} \left\langle [g(|x|/t) - 1 + t^{-1}|x|g'(|x|/t)]Q_0\chi_0(\mathbf{E}(t),\mathbf{H}(t)), Q_0\chi_0(\mathbf{E}(t),\mathbf{H}(t))\right\rangle_{L^2(\mathbb{R}^3)}$$

$$-t^{-1}\left\langle (2\tilde{S} + t^{-1}|x|S)g'(|x|/t)Q_0\chi_0(\mathbf{E}(t),\mathbf{H}(t)), Q_0\chi_0(\mathbf{E}(t),\mathbf{H}(t))\right\rangle_{L^2(\mathbb{R}^3)}.$$

In the following estimates C_j are constants independent of s. Lemma 4.9 and (4.41) yield by Hölder's inequality and assumption (4.9)

$$|h_1(s)| \le C_1\|(1+|x|)^{-1/2}Sg(|x|/s)Q_0\chi_0(\mathbf{E}(s),\mathbf{H}(s))\|_{H^1(\mathbb{R}^3)}$$

$$\|(1+|x|)^{1/2}\chi_0\mathcal{R}\partial_s\mathbf{P}(s)\|_{L^{q_0}(\mathbb{R}^3)}$$

$$\le C_1\|(1+|x|)^{-1/2}Sg(|x|/s)Q_0\chi_0(\mathbf{E}(s),\mathbf{H}(s))\|_{H^1(\mathbb{R}^3)}$$

$$\|(1+|x|)^{1/2}\gamma^{1/2}\|_{L^{2r_0}(\mathbb{R}^3)}\|\gamma^{-1/2}\partial_s\mathbf{P}(s)\|_{L^2(G)}$$

$$\le C_2 s^{1/2}\|\partial_s\mathbf{P}(s)\|_{\mathcal{G}}.$$

For all $T > 0$ one obtains

$$t^{-1}\int_1^t |h_1(s)|ds \le t^{-1}\int_1^T |h_1(s)|ds + C_2 t^{-1}\int_T^t s^{1/2}\|\partial_s\mathbf{P}(s)\|_{\mathcal{G}}ds$$

$$\le t^{-1}\int_1^T |h_1(s)|ds + C_2\left(\int_T^t \|\partial_s\mathbf{P}(s)\|_{\mathcal{G}}^2 ds\right)^{1/2}$$

and hence by Lemma 2.2

$$\limsup_{t\to\infty} t^{-1}\int_1^t |h_1(s)|ds \le C_2\left(\int_T^\infty \|\partial_s\mathbf{P}(s)\|_{\mathcal{G}}^2 ds\right)^{1/2}$$

for all $T > 0$, which implies that

$$t^{-1}\int_1^t |h_1(s)|ds \stackrel{t\to\infty}{\longrightarrow} 0. \tag{4.51}$$

Next

$$|h_0(t)| \le C_3\|(1+|x|)^{-1}Sg(|x|/t)Q_0\chi_0(\mathbf{E}(t),\mathbf{H}(t))\|_{H^1}\|(1+|x|)\chi_0(\mathbf{j}(t),0)\|_{L^{q_0}}$$

$$\le C_4\|\mathbf{j}(t)\|_{L^2(B_{R_1})} \le C_4\|\mathbf{j}(t)\|_{L^2(\Omega)}$$

by assumption (4.10) which implies that

$$t^{-1}\int_0^t |h_0(s)|ds \overset{t\to\infty}{\longrightarrow} 0. \tag{4.52}$$

Similarly

$$|h_2(t)| \leq C_5\|(1+|x|)^{-1}Sg(|x|/t)Q_0\chi_0(\mathbf{E}(t),\mathbf{H}(t))\|_{H^1}\|(1+|x|)\mathcal{C}_0(\mathbf{E}(t),\mathbf{H}(t))\|_{L^{q_0}}$$

$$\leq C_6\|(\mathbf{E}(t),\mathbf{H}(t))\|_{L^2(B_{R_0})}$$

and hence by Corollary 4.8

$$t^{-1}\int_0^t |h_2(s)|ds \overset{t\to\infty}{\longrightarrow} 0. \tag{4.53}$$

Since $g'(|x|/t) = 0$ and $g(|x|/t) = 1$ if $|x| \leq at$ with some $a > 1$, Lemma 4.10(iii) yields

$$t^{-1}\int_0^t |h_3(s)|ds \leq C_7 t^{-1}\int_0^t \|Q_0\chi_0(\mathbf{E}(s),\mathbf{H}(s))\|_{L^2(\{|x|\geq as\})}ds \overset{t\to\infty}{\longrightarrow} 0. \tag{4.54}$$

Now, it follows from (4.49)–(4.54) and Lemma 4.10 that

$$\lim_{t\to\infty} t^{-1}F(t) = \lim_{t\to\infty} t^{-1}\int_1^t F'(s)ds$$

$$= \lim_{t\to\infty} t^{-1}\int_1^t \|Q_0\chi_0(\mathbf{E}(s),\mathbf{H}(s))\|^2_{L^2(\mathbb{R}^3)}ds = \mathcal{E}_\infty.$$

This completes the proof. □

Now the main results of this section, (1.7) and (1.9), can be proved.

THEOREM 4.12. *For all $a < 1$ and $b > 1$ one has*

$$\|(\mathbf{E}(t),\mathbf{H}(t))\|_{L^2(\Omega\cap B_{at})} \overset{t\to\infty}{\longrightarrow} 0 \tag{4.55}$$

and

$$\|(\mathbf{E}(t),\mathbf{H}(t)) - t^{-1}S\chi_{\{at\leq|x|\leq bt\}}(\mathbf{E}(t),\mathbf{H}(t))\|_X \overset{t\to\infty}{\longrightarrow} 0.$$

Furthermore

$$\|(1-Q_0)\chi_0(\mathbf{E}(t),\mathbf{H}(t))\|_{L^2(\mathbb{R}^3)} \overset{t\to\infty}{\longrightarrow} 0.$$

Proof. Suppose $\delta > 0$. Choose $g \in C_0^\infty(\mathbb{R},[0,\infty))$ with $g(y) = 1$ on $[0,1+\delta/2]$ and $g(u) = 0$ for all $u > 1+\delta$. Then

$$\|(1-Q_0)\chi_0(\mathbf{E}(t),\mathbf{H}(t))\|^2_{L^2(\mathbb{R}^3)}$$

$$= \|\chi_0(\mathbf{E}(t),\mathbf{H}(t))\|^2_{L^2(\mathbb{R}^3)} - \|Q_0\chi_0(\mathbf{E}(t),\mathbf{H}(t))\|^2_{L^2(\mathbb{R}^3)}$$

$$\leq \|(\mathbf{E}(t),\mathbf{H}(t))\|_X^2 - \|Q_0\chi_0(\mathbf{E}(t),\mathbf{H}(t))\|_{L^2(\mathbb{R}^3)}^2$$

$$\leq \|(\mathbf{E}(t),\mathbf{H}(t))\|_X^2 - (1+\delta)^{-1}t^{-1}$$

$$\langle Sg(|x|/t)Q_0\chi_0(\mathbf{E}(t),\mathbf{H}(t)), Q_0\chi_0(\mathbf{E}(t),\mathbf{H}(t))\rangle_{L^2(\mathbb{R}^3)}.$$

Theorem 4.11 yields

$$\limsup_{t\to\infty} \|(1-Q_0)\chi_0(\mathbf{E}(t),\mathbf{H}(t))\|_{L^2(\mathbb{R}^3)}^2 \leq \left(1-(1+\delta)^{-1}\right)\mathcal{E}_\infty,$$

since $\limsup_{t\to\infty} \|(\mathbf{E}(t),\mathbf{H}(t))\|_X^2 \leq \mathcal{E}_\infty$ by (2.23). By letting $\delta \to 0$ this implies

$$\lim_{t\to\infty} \|(1-Q_0)\chi_0(\mathbf{E}(t),\mathbf{H}(t))\|_{L^2(\mathbb{R}^3)}^2 = 0. \tag{4.56}$$

This improves assertion (i) of Lemma 4.10. Next, one obtains from Theorem 4.2, the boundedness of supp $(1-\chi_0)$, Lemma 2.2(i), (4.56), and Theorem 4.11 that for all $\beta > 1$

$$\lim_{t\to\infty} t^{-1}\langle S\chi_{\{|x|\leq\beta t\}}(\mathbf{E}(t),\mathbf{H}(t)), (\mathbf{E}(t),\mathbf{H}(t))\rangle_X \tag{4.57}$$

$$= \lim_{t\to\infty} t^{-1}\langle Sg(|x|/t)(\mathbf{E}(t),\mathbf{H}(t)), (\mathbf{E}(t),\mathbf{H}(t))\rangle_X$$

$$= \lim_{t\to\infty} t^{-1}\langle Sg(|x|/t)\chi_0(\mathbf{E}(t),\mathbf{H}(t)), \chi_0(\mathbf{E}(t),\mathbf{H}(t))\rangle_{L^2(\mathbb{R}^3)}$$

$$= \lim_{t\to\infty} t^{-1}\langle Sg(|x|/t)Q_0\chi_0(\mathbf{E}(t),\mathbf{H}(t)), Q_0\chi_0(\mathbf{E}(t),\mathbf{H}(t))\rangle_{L^2(\mathbb{R}^3)} = \mathcal{E}_\infty.$$

Here a function $g \in C_0^\infty(\mathbb{R},[0,\infty))$ with the properties $g(y) = 1$ on $[0,\beta]$ and $g(u) = 0$ for all $u > 2\beta$ is chosen.

Let $\beta > 1$. Then one obtains from (4.57)

$$\int_{\Omega\cap B_{at}} |(\mathbf{E}(t),\mathbf{H}(t))|^2 dx \leq \int_{\Omega\cap B_{\beta t}} |(\mathbf{E}(t),\mathbf{H}(t))|^2 dx$$

$$-\beta^{-1}t^{-1}\int_{\{at\leq|x|\leq\beta t\}} |x| \quad |(\mathbf{E}(t),\mathbf{H}(t))|^2 dx$$

$$\leq \|(\mathbf{E}(t),\mathbf{H}(t))\|_X^2 - \beta^{-1}t^{-1}\int_{\Omega\cap B_{\beta t}} |x| \quad |(\mathbf{E}(t),\mathbf{H}(t))|^2 dx$$

$$+\beta^{-1}a\int_{\Omega\cap B_{at}} |(\mathbf{E}(t),\mathbf{H}(t))|^2 dx$$

$$\leq \|(\mathbf{E}(t),\mathbf{H}(t))\|_X^2 - \beta^{-1}t^{-1}\langle S\chi_{\{|x|\leq\beta t\}}(\mathbf{E}(t),\mathbf{H}(t)), (\mathbf{E}(t),\mathbf{H}(t))\rangle_X$$

$$+a\int_{\Omega\cap B_{at}}|(\mathbf{E}(t),\mathbf{H}(t))|^2dx.$$

Hence

$$(1-a)\int_{\Omega\cap B_{at}}|(\mathbf{E}(t),\mathbf{H}(t))|^2dx\le\|(\mathbf{E}(t),\mathbf{H}(t))\|_X^2$$

$$-\beta^{-1}t^{-1}\langle S\chi_{\{|x|\le\beta t\}}(\mathbf{E}(t),\mathbf{H}(t)),(\mathbf{E}(t),\mathbf{H}(t))\rangle_X.$$

Invoking (4.57) one gets

$$(1-a)\limsup_{t\to\infty}\int_{\Omega\cap B_{at}}|(\mathbf{E}(t),\mathbf{H}(t))|^2dx\le(1-\beta^{-1})\mathcal{E}_\infty\text{ for all }\beta>1.$$

By letting $\beta\to1$ this impliess

$$\|(\mathbf{E}(t),\mathbf{H}(t))\|_{L^2(\Omega\cap B_{at})}\overset{t\to\infty}{\longrightarrow}0. \tag{4.58}$$

This completes the proof of the first assertion (4.55).

Suppose $\beta>1$. Then it follows from Theorem 4.2 that

$$\limsup_{t\to\infty}t^{-1}\|S\chi_{\{|x|\le bt\}}(\mathbf{E}(t),\mathbf{H}(t))\|_X\le\limsup_{t\to\infty}t^{-1}\|S\chi_{\{|x|\le\beta t\}}(\mathbf{E}(t),\mathbf{H}(t))\|_X$$

$$\le\beta\limsup_{t\to\infty}\|(\mathbf{E}(t),\mathbf{H}(t))\|_X\le\beta\mathcal{E}_\infty^{1/2}.$$

Letting $\beta\to1$ this yields

$$\limsup_{t\to\infty}t^{-1}\|S\chi_{\{|x|\le bt\}}(\mathbf{E}(t),\mathbf{H}(t))\|_X\le\mathcal{E}_\infty^{1/2}.$$

By (4.57) one obtains

$$\limsup_{t\to\infty}\|t^{-1}S\chi_{\{|x|\le bt\}}(\mathbf{E}(t),\mathbf{H}(t))-(\mathbf{E}(t),\mathbf{H}(t))\|_X^2$$

$$=\limsup_{t\to\infty}\big(t^{-2}\|S\chi_{\{|x|\le bt\}}(\mathbf{E}(t),\mathbf{H}(t))\|_X^2$$

$$-2t^{-1}\left\langle S\chi_{\{|x|\le bt\}}(\mathbf{E}(t),\mathbf{H}(t)),(\mathbf{E}(t),\mathbf{H}(t))\right\rangle_X+\|(\mathbf{E}(t),\mathbf{H}(t))\|_X^2)\le0,$$

which completes the proof. □

REMARK 2. *The above theorem does not provide any information on the asymptotic behavior of* $\mathbf{P}$. *But if the potential is quadratically coercive in the sense that*

$$\mathbf{p}(\nabla_PV)(x,\mathbf{p})\ge a_0|\mathbf{p}|^2\text{ for all }\mathbf{p}\in\mathbb{R}^3$$

with some $a_0>0$, *it follows easily from a similar estimate as* (2.17) *that*

$$\|\mathbf{P}(t)\|_{L^2(G\cap B_R)}\overset{t\to\infty}{\longrightarrow}0\text{ for all }R>0 \tag{4.59}$$

provided that $\mathbf{E}$ *satisfies*

$$\|\mathbf{E}(t)\|_{L^2(\Omega\cap B_R)} \overset{t\to\infty}{\longrightarrow} 0 \text{ for all } R>0. \tag{4.60}$$

In particular, (4.59) *holds if condition* (1.8) *is fulfilled by Theorem* 4.12. *Furthermore, it turns out that condition* (1.8) *is also necessary for the local decay of the electromagnetic field in this case. This can be seen as follows. If* $\|(\mathbf{E}(t),\mathbf{H}(t))\|_{L^2(\Omega\cap B_R)} \overset{t\to\infty}{\longrightarrow} 0$ *for all* $R>0$, *then also* (4.59) *holds and therefore*

$$(\mathbf{E}(t),\mathbf{H}(t)) \overset{t\to\infty}{\longrightarrow} 0 \text{ in } X \text{ weakly} \tag{4.61}$$

and

$$\mathbf{P}(t) \overset{t\to\infty}{\longrightarrow} 0 \text{ in } L^2(G) \text{ weakly.} \tag{4.62}$$

Hence one obtains from (4.1), (4.61), *and* (4.62) *by letting* $t\to\infty$ *that*

$$(1-Q)(\mathbf{D}_1,\mathbf{H}_0) = (1-Q)\left((\mathbf{E}_0,\mathbf{H}_0)+\mathcal{R}\mathbf{P}(0)-\int_0^\infty(\mathbf{j}(s),0)ds\right)$$

$$= w-\lim_{t\to\infty}(1-Q)\left((\mathbf{E}(t),\mathbf{H}(t))+\mathcal{R}\mathbf{P}(t)\right)=0,$$

whence (1.8).

Invoking a result in [9] concerning the linear inhomogeneous Maxwell equations without polarization it can be shown that the solution $(\mathbf{E},\mathbf{H})$ of (1.1)–(1.5) is asymptotically free in the sense that there exists a uniquely determined pair of functions $(\mathbf{F}_0,\mathbf{G}_0)\in L^2(\mathbb{R}^3)$ with $\operatorname{div}\mathbf{F}_0 = \operatorname{div}\mathbf{G}_0 = 0$ such that

$$\|(\mathbf{E}(t),\mathbf{H}(t))-(\mathbf{F}(t),\mathbf{G}(t))\|_{L^2(\Omega)} \overset{t\to\infty}{\longrightarrow} 0. \tag{4.63}$$

Here $(\mathbf{F},\mathbf{G})\in C(\mathbb{R},L^2(\mathbb{R}^3,\mathbb{C}^6))$ denotes the solution to Maxwell's equations in the whole space, that is

$$\partial_t\mathbf{F} = \operatorname{curl}\mathbf{G},\quad \partial_t\mathbf{G} = -\operatorname{curl}\mathbf{F}, \tag{4.64}$$

supplemented by the initial-condition

$$\mathbf{F}(0,x)=\mathbf{F}_0(x),\mathbf{G}(0,x)=\mathbf{G}_0(x). \tag{4.65}$$

This means that the solution to (1.1)–(1.4) behaves asymptotically like a free space solution to (4.64), (4.65) as $t\to\infty$. In what follows suppose that in addition

$$(1+|x|)^{1+\alpha_0}\gamma^{1/2}\in L^{r_1}(G) \tag{4.66}$$

for some $\alpha_0>0$ and $r_1\in(3,\infty)$. This condition is obviously fulfilled in the case where the set G is bounded.

THEOREM 4.13. *The strong limit*

$$\mathbf{U}\overset{\text{def}}{=}\lim_{t\to\infty}\exp(-tB_0)J^*(\mathbf{E}(t),\mathbf{H}(t))$$

exists in $L^2(\mathbb{R}^3)$ *and* $\mathbf{U}\in(\ker B_0)^\perp$. *Here* $J^*:L^2(\Omega)\to L^2(\mathbb{R}^3)$ *provides the extension by zero on* $\mathbb{R}^3\setminus\Omega$.

Proof. It follows from Theorem 4.12 that for all $a < 1 < b$

$$\lim_{t\to\infty} \|t^{-1} S\chi_{\{at\le|x|\le bt\}} Q_0\chi_0(\mathbf{E}(t),\mathbf{H}(t)) - J^*(\mathbf{E}(t),\mathbf{H}(t))\|_{L^2} = 0. \tag{4.67}$$

Let g be defined as in [9, Theorem 8] by $g(t,u) \stackrel{\text{def}}{=} c_\alpha t^{-1-\alpha} u^\alpha$ for $u \le (1+\alpha)^{-1}\alpha t$ and $g(t,u) \stackrel{\text{def}}{=} u^{-1}$ for $u \ge (1+\alpha)^{-1}\alpha t$. Here $\alpha \stackrel{\text{def}}{=} \alpha_0/2 > 0$ with α_0 as in assumption (4.66) and $c_\alpha \stackrel{\text{def}}{=} (1+\alpha^{-1})^{1+\alpha}$.

Since $g(t,t) = t^{-1}$, it follows easily from (4.67), Theorem 4.2, and Theorem 4.12 that

$$\lim_{t\to\infty} \|Sg(t,|x|) Q_0\chi_0(\mathbf{E}(t),\mathbf{H}(t)) - J^*(\mathbf{E}(t),\mathbf{H}(t))\|_{L^2} = 0. \tag{4.68}$$

Next Theorem 4.12 yields further $\lim_{t\to\infty} \|(1-Q_0)J^*(\mathbf{E}(t),\mathbf{H}(t))\|_{L^2} = 0$, and hence by (4.68)

$$\lim_{t\to\infty} \|L(t)\chi_0(\mathbf{E}(t),\mathbf{H}(t)) - J^*(\mathbf{E}(t),\mathbf{H}(t))\|_{L^2} = 0, \tag{4.69}$$

where $L(t) \stackrel{\text{def}}{=} Q_0 Sg(t,|x|)Q_0 \in B(L^2,L^2)$ with g defined as above.

The following result concerning the inhomogeneous linear Maxwell equations can be found in [9, Theorem 8].

THEOREM 4.14. *Suppose that* $\mathbf{u} \in L^\infty((0,\infty),D(B)) \cap C([0,\infty),X)$ *solves* $\partial_t\mathbf{u} = B\mathbf{u} + \mathbf{f}$, *where* $\mathbf{f} \in L^1_{loc}([0,\infty),X)$ *obeys* $(1+|x|)^{1+\alpha_0}\mathbf{f} \in L^1((0,\infty),L^{q_1}(\Omega)) + L^\infty((0,\infty),L^{q_1}(\Omega))$.

Then the strong limit

$$\lim_{t\to\infty} \exp(-tB_0)L(t)\chi_0\mathbf{u}(t)$$

with respect to the $L^2(\mathbb{R}^3)$*-topology exists.*

Here $q_1 \in [6/5,2)$ is defined by $1/q_1 = 1/2 + 1/r_1$, where $\alpha_0 > 0$ and $r_1 \in [3,\infty)$ are as in assumption (4.66).

In order to apply Theorem 4.14, let $\mathbf{u}(t) \stackrel{\text{def}}{=} (\mathbf{E}(t),\mathbf{H}(t))$ and $\mathbf{f}(t) \stackrel{\text{def}}{=} \partial_t\mathcal{R}\mathbf{P}(t) + (\mathbf{j}(t),0)$. With the assumptions (4.66), (4.10), and Lemma 2.2 one has $(1+|x|)^{1+\alpha_0}\mathbf{f} \in L^\infty((0,\infty),L^{q_1}(\Omega))$. Hence $\mathbf{u}$ satisfies the conditions of Theorem 4.14, which implies the existence of the limit

$$\lim_{t\to\infty} \exp(-tB_0)L(t)\chi_0(\mathbf{E}(t),\mathbf{H}(t)). \tag{4.70}$$

By (4.69) one obtains the existence of the limit

$$\mathbf{U} \stackrel{\text{def}}{=} \lim_{t\to\infty} \exp(-tB_0)J^*(\mathbf{E}(t),\mathbf{H}(t)) \text{ in } L^2(\mathbb{R}^3). \tag{4.71}$$

Since ran $L(t) \subset$ ran Q_0, it follows from (4.69) and (4.71) that $\mathbf{U} \in \overline{\text{ran } B_0} = (\ker B_0)^\perp$, i.e., div $(\underline{\mathbf{U}}_j) = 0$ on $\mathbb{R}^3$. Now, it follows easily that $(\mathbf{F},\mathbf{G}) \stackrel{\text{def}}{=} \exp(tB_0)\mathbf{U}$ satisfies (4.63). ∎

Acknowledgments. The author thanks the anonymous referees for reading the manuscript carefully and for their comments, which were helpful in improving the presentation of the results.

REFERENCES

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

[2] H. Barucq and B. Hanouzet, *Asymptotic behavior of solutions to Maxwell's equations in bounded domains with absorbing Silver-Müller's condition on the exterior boundary*, Asympt. Anal., 15 (1997), pp. 25–40.

[3] R. Boyd, *Nonlinear Optics*, Academic Press, New York, 1992.

[4] G. Carbou and P. Fabrie, *Time average in micromagnetism*, J. Differential Equations, 147 (1998), pp. 383–409.

[5] C. M. Dafermos, *Asymptotic behavior of solutions of evolution equations*, in Nonlinear Evolution Equations, Academic Press, New York, 1978, pp. 103–123.

[6] A. Haraux, *Stabilization of trajectories for some weakly damped hyperbolic equations*, J. Differential Equations, 59 (1985), pp. 145–154.

[7] F. Jochmann, *A compactness result for vector fields with divergence and curl in $L^q(\Omega)$ involving mixed boundary conditions*, Appl. Anal., 66 (1997), pp. 198–203.

[8] F. Jochmann, *The semistatic limit for Maxwell's equations in an exterior domain*, Comm. Partial Differential Equations, 23 (1998), pp. 2035–2076.

[9] F. Jochmann, *Asymptotic behaviour of solutions to Maxwell's equations in exterior domains*, Asympt. Anal., 21 (1999), pp. 331–363.

[10] F. Jochmann, *Regularity of weak solutions to Maxwell's equations with mixed boundary conditions*, Math. Methods Appl. Sci., 22 (1999), pp. 1255–1274.

[11] F. Jochmann, *Asymptotic behaviour of solutions to a class of semilinear hyperbolic systems in arbitrary domains*, J. Differential Equations, 160 (2000), pp. 439–466.

[12] J. L. Joly, G. Métivier, and J. Rauch, *Global solvability of the anharmonic oscillator model from nonlinear optics*, SIAM J. Math. Anal., 27 (1996), pp. 905–913.

[13] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer, New York, 1983.

[14] R. Picard, *An elementary proof for a compact embedding result in generalized electromagnetic theory*, Math. Z., 187 (1984), pp. 151–161.

[15] M. Slemrod, *Weak asymptotic decay via a relaxed invariance principle for a wave equation with nonlinear nonmonotone damping*, Proc. Roy. Soc. Edinburgh Sect. A, 113 (1989), pp. 87–97.

[16] C. Weber, *A local compactness theorem for Maxwell's equations*, Math. Methods Appl. Sci., 2 (1980), pp. 12–25.

SIAM J. MATH. ANAL.
Vol. 32, No. 4, pp. 916–927

ON A SINGULAR NONLINEAR DIRICHLET PROBLEM WITH A CONVECTION TERM*

ZHIJUN ZHANG† AND JIANNING YU‡

Abstract. We consider a singular nonlinear Dirichlet problem with a convection term. The approach is based on existence regularity theory and a subsolution-supersolution method. Nonexistence and regularity results are also given.

Key words. semilinear elliptic equation, singular nonlinearity, Dirichlet problem, positive solution, subsolution-supersolution method

AMS subject classifications. 35J65, 35B05, 35075, 35R05

PII. S0036141097332165

1. Introduction and main results.

Let Ω be a bounded domain in R^N ($2 \leq N$) with $C^{2+\beta}$ boundary $\partial\Omega$ for some $\beta \in (0,1)$. Consider the boundary value problem

$$\begin{cases} -\Delta u = \dfrac{1}{u^\alpha} + \lambda |\nabla u|^p + \sigma, \quad 0 < u \quad \text{in } \Omega, \\ u|_{\partial\Omega} = 0, \end{cases} \tag{1.1}$$

where $0 < \alpha$, $0 < p \leq 2$, $0 \leq \lambda$, and $0 \leq \sigma$. This problem arises in certain problems in fluid mechanics and pseudoplastic flow (see [2, 16]).

As is well known, the model problem

$$\begin{cases} -\Delta u = \dfrac{1}{u^\alpha} + \lambda u^q, \quad 0 < u \quad \text{in } \Omega, \\ u|_{\partial\Omega} = 0, \end{cases} \tag{1.2}$$

was discussed and extended to the cases of more general problems in a number of works; see, for instance, [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]. It is shown that if $0 < q < 1$, then (1.2) has at least one classical solution in $C^{2+\beta}(\Omega) \bigcap C(\overline{\Omega})$ for all $\lambda \geq 0$, and if $q \geq 1$, then there exists $\overline{\lambda} \in (0, \infty)$ such that (1.2) has at least one classical solution for $\lambda \in [0, \overline{\lambda})$, and (1.2) has no classical solution for $\lambda > \overline{\lambda}$. In particular, if $\lambda = 0$, it is shown that (1.2) has a unique classical solution u. Moreover, u has the following properties (see [3, 7]):

$(1)^0$ if $\alpha > 1$, then u is not in $C^1(\overline{\Omega})$;

$(2)^0$ if $\alpha < 3$, then $u \in H_0^1(\Omega)$;

$(3)^0$ if $\alpha \geq 3$, then u is not in $H_0^1(\Omega)$.

It is worth noting that the classical solution of (1.2) with $\lambda = 0$ is not a weak solution for $\alpha \geq 3$. This is very different from that in the case of nonsingular Dirichlet problems; see, for instance, [13, 14, 15, 18].

*Received by the editors December 30, 1997; accepted for publication (in revised form) July 27, 2000; published electronically December 28, 2000. This work was supported in part by the National Natural Science Foundation of the People's Republic of China.

http://www.siam.org/journals/sima/32-4/33216.html

†Department of Mathematics and Information Science, Yantai University, Yantai 264005, Shandong, People's Republic of China (math@ytu.edu.cn).

‡Department of Basic Courses, Lanzhou Railway Institute, Lanzhou 730070, People's Republic of China (yujn@lzri.edu.cn).

For the problem (1.1), only partial results are known. In [8], it is shown that (1.1) has just one classical solution for $0 < \alpha$, $0 < p < 1$, $0 \le \lambda$, and $0 \le \sigma$. Other results are not known for the problem (1.1).

In this paper, among other things, we have studied the problem (1.1) for $0 < p \le 2$, $0 \le \lambda$, and $0 \le \sigma$. Our main results are the following.

THEOREM 1.1. *If $p = 2$, then (1.1) has a unique classical solution for $\lambda\sigma < \lambda_1$ and no solution in $H_0^1(\Omega)$ for $\lambda\sigma \ge \lambda_1$, where λ_1 is the first eigenvalue for the Laplacian with Dirichlet boundary condition.*

THEOREM 1.2. *If $0 < p < 2$, then there exists $\overline{\lambda}(p, \sigma) \in (0, \infty]$ such that (1.1) has a unique classical solution for $\lambda \in [0, \overline{\lambda})$ and no classical solution for $\lambda > \overline{\lambda}$. In addition,*

(i) *if $1 < p < 2$ and $0 < \sigma$, then $\overline{\lambda} < \infty$;*

(ii) *if $1 < p < 2$ and $\sigma = 0$ or $0 < p < 1$, then $\overline{\lambda} = \infty$.*

THEOREM 1.3. *The unique classical solution u of (1.1) has the following properties:*

1^0 *if $\alpha > 1$, then u is not in $C^1(\overline{\Omega})$.*

2^0 *if $\alpha < 3$, then $u \in H_0^1(\Omega)$.*

3^0 *if $\alpha \ge 3$ and one of the following conditions is satisfied*

$$(c_1)\ p = 2, \quad (c_2)\ 0 < p < 1, \quad (c_3)\ 1 < p < 2 \text{ and } \sigma = 0,$$

then u is not in $H_0^1(\Omega)$.

Remark 1.4. It is not known, if $p = 1$, whether $\overline{\lambda}$ in Theorem 1.2 is finite or not, and if $1 < p < 2$ and $0 < \sigma$, or $p = 1$, whether u in Theorem 1.3 has the property 3^0 or not.

Remark 1.5. It is obvious that (1.1) has no solution in $C^2(\overline{\Omega})$.

Remark 1.6. The problem (1.1) or (1.2) is significantly different from that problem which has been studied in [11, 12, 20, 21, 22].

The outline of this article is as follows. In section 2 we first make an attempt to develop the famous existence theorem of Amann in [13] into the more general singular boundary value problem and give some preliminary conditions. Then we give the proof of our main results. Our proof is based on the results in [3, 5, 7, 8, 9, 22, 23] and is made possible by the choice of supersolutions and subsolutions.

By a solution of (1.1) we mean, until further notice, a classical solution, i.e., a function belonging to $C^{2+\beta}(\Omega) \bigcap C(\overline{\Omega})$ and satisfying (1.1).

2. Preliminaries. First we give some preliminary considerations and lemmas.

Let ϕ_1 denote a positive eigenfunction corresponding to the first eigenvalue λ_1, and e denote the unique solution of the problem (see [14, 15])

$$\begin{cases} -\Delta u = 1 & \text{in } \Omega, \\ u|_{\partial\Omega} = 0. \end{cases} \tag{2.1}$$

As is well known, $e, \phi_1 \in C^{2+\beta}(\overline{\Omega})$, $0 < e$, $0 < \phi_1$ in Ω, and $\frac{\partial e(x)}{\partial \vec{n}} < 0$, $\frac{\partial \phi_1(x)}{\partial \vec{n}} < 0$ for all $x \in \partial\Omega$, where $\vec{n}$ denotes the outward normal to $\partial\Omega$.

Put

$$|u|_\infty = \max\{|u(x)| \,|\, x \in \overline{\Omega},\ u \in C(\overline{\Omega})\}.$$

We can obtain the following.

LEMMA 2.1. *There exist positive constants C_1 and C_2 such that*

$$C_1\phi_1 \le e \le C_2\phi_1 \qquad \text{on } \overline{\Omega}.$$

LEMMA 2.2. $\int_\Omega [\frac{1}{\phi_1(x)}]^s\,dx < \infty$ *if and only if* $s < 1$.

LEMMA 2.3 (see [7]). *Let ω be the unique solution of the problem*

$$\begin{cases} -\Delta u = \dfrac{1}{u^\alpha}, \quad 0 < u \quad \text{in } \Omega, \\ u|_{\partial\Omega} = 0. \end{cases} \tag{2.2}$$

Then if $\alpha > 1$, there exist positive constants C_3 and C_4 such that

$$C_3\phi_1^{2/(1+\alpha)} \le \omega \le C_4\phi_1^{2/(1+\alpha)} \qquad \text{on } \overline{\Omega}.$$

LEMMA 2.4 (see [17, Theorem 2.1]). *There exists $\lambda^* < \infty$ such that the following problem has a solution neither in $W^{1,1}(\Omega)$ nor in $C(\overline{\Omega})\bigcap C^2(\Omega)$ for $\lambda > \lambda^*$*

$$\begin{cases} -\Delta u = |\nabla u|^p + \lambda \quad \text{in } \Omega, \\ u|_{\partial\Omega} = 0, \end{cases} \tag{2.3}$$

where $1 < p$ and $0 < \lambda$.

Now we consider the more general problem

$$\begin{cases} -\Delta u = g(u) + h(u) + k(\nabla u), \quad 0 < u \quad \text{in } \Omega, \\ u|_{\partial\Omega} = 0, \end{cases} \tag{2.4}$$

where the functions $g(s)$, $h(s)$, and $k(\eta)$ satisfy the following conditions:

(I_1) $h \in C^1([0,\infty), [0,\infty))$;
(I_2) $g \in C^1((0,\infty), [0,\infty))$;
(I_3) $\lim_{s\to 0+} g(s) = \infty$;
(I_4) $g'(s) \le 0\ \forall s > 0$, i.e., g is nonincreasing in $(0,\infty)$;
(I_5) $k \in C^1(R^N, [0,\infty))$;
(I_6) there exists a positive constant C_5 such that

$$k(\eta) \le C_5(1 + |\eta|^2) \qquad \forall \eta \in R^N.$$

DEFINITION 2.5. *A function $\underline{u}$ is called a subsolution of* (2.4) *if $\underline{u} \in C(\overline{\Omega})\bigcap C^2(\Omega)$ and*

$$\begin{cases} -\Delta \underline{u} \le g(\underline{u}) + h(\underline{u}) + k(\nabla \underline{u}), \quad 0 < \underline{u} \quad \text{in } \Omega, \\ \underline{u}|_{\partial\Omega} = 0. \end{cases} \tag{2.5}$$

DEFINITION 2.6. *A function $\overline{u}$ is called a supersolution of* (2.4) *if $\overline{u} \in C(\overline{\Omega})\bigcap C^2(\Omega)$ and*

$$\begin{cases} -\Delta \overline{u} \ge g(\overline{u}) + h(\overline{u}) + k(\nabla \overline{u}), \quad 0 < \overline{u} \quad \text{in } \Omega, \\ \overline{u}|_{\partial\Omega} = 0. \end{cases} \tag{2.6}$$

LEMMA 2.7 (see [3]). *Under the conditions (I_2)–(I_4), the problem*

$$\begin{cases} -\Delta u = g(u + \varepsilon), \quad 0 < u \quad \text{in } \Omega, \\ u|_{\partial\Omega} = 0 \end{cases} \tag{2.7}$$

has a unique solution $\omega_\varepsilon \in C^{2+\beta}(\overline{\Omega})$ *for any sufficiently small positive number* ε. *Moreover, for any* $C^{2+\beta}$*-smooth domain* $\Omega' \subset\subset \Omega$, $\{\omega_\varepsilon\}$ *has a subsequence which converges uniformly in the* $C^2(\overline{\Omega'})$ *norm to the unique solution* $\omega \in C^{2+\beta}(\Omega) \bigcap C(\overline{\Omega})$ *of the following problem:*

$$\begin{cases} -\Delta u = g(u), \quad 0 < u \quad \text{in } \Omega, \\ u|_{\partial\Omega} = 0, \end{cases} \tag{2.8}$$

and $\omega_\varepsilon \le \omega \le \omega_\varepsilon + \varepsilon$ *on* $\overline{\Omega}$.

Obviously, under the conditions (I_2)–(I_6), ω is a subsolution of (2.4).

Basic to our subsequent discussions is the following theorem which is formulated in terms of supersolution and subsolution.

LEMMA 2.8. *If* (2.4) *has a supersolution* $\overline{u}$ *such that* $\omega \le \overline{u}$ *in* Ω, *then* (2.4) *has at least one solution* $u \in C^{2+\beta}(\Omega) \bigcap C(\overline{\Omega})$ *satisfying*

$$\omega \le u \le \overline{u} \qquad \text{on } \overline{\Omega}.$$

Proof. Consider the perturbed problem of (2.4)

$$\begin{cases} -\Delta u = g(u+\varepsilon) + h(u) + k(\nabla u), \quad 0 < u \quad \text{in } \Omega, \\ u|_{\partial\Omega} = 0, \end{cases} \tag{2.9}$$

where ε is a sufficiently small positive number.

Clearly ω_ε is a subsolution of (2.9) and $\overline{u}$ is supersolution of (2.9).

In order to prove the existence of solutions of (2.9), we can easily prove that there exists a function $f_\varepsilon(s,\eta) \in C^1(R \times R^N)$ such that

(I_7) $f_\varepsilon(s,\eta) = g(s+\varepsilon) + h(s) + k(\eta)$ $\forall (s,\eta) \in [0,\infty) \times R^N$,

(I_8) $|f_\varepsilon(s,\eta)| \le C_\varepsilon(|s|)(1+|\eta|^2)$ $\forall (s,\eta) \in R \times R^N$,

where $C_\varepsilon \in C^1([0,\infty), [0,\infty))$ and is a nondecreasing function.

Now we consider the following perturbed problem:

$$\begin{cases} -\Delta u = f_\varepsilon(u, \nabla u), \quad 0 < u \quad \text{in } \Omega, \\ u|_{\partial\Omega} = 0. \end{cases} \tag{2.10}$$

Since $\overline{u} \ge \omega \ge \omega_\varepsilon > 0$ in Ω, we know that ω_ε is also a subsolution of (2.10). It follows by the first theorem of Amann [13] that (2.10) has a maximal solution $u_\varepsilon \in C^{2+\beta}(\overline{\Omega})$ in order interval $[\omega_\varepsilon, \overline{u}]$. Thus u_ε is also a solution of (2.9).

Now we need to estimate $\{u_\varepsilon\}$. For any $C^{2+\beta}$-smooth domain $\Omega' \subset\subset \Omega$, take Ω_i, $i = 1,2,3$, with $C^{2+\beta}$-smooth boundaries such that

$$\Omega' \subset\subset \Omega_1 \subset\subset \Omega_2 \subset\subset \Omega_3 \subset\subset \Omega.$$

Let

$$\overline{f}_\varepsilon(x) = f_\varepsilon(u_\varepsilon(x), \nabla u_\varepsilon(x)), \qquad x \in \overline{\Omega}_3.$$

Since $-\Delta u_\varepsilon = \overline{f}_\varepsilon(x)$ on $\overline{\Omega}_3$, by the interior estimate theorem of Ladyzenskaya and Ural'ceva (see [14, Theorem 3.1, p. 266]), we get a positive constant C_6 independent of ε such that

$$\max_{x \in \overline{\Omega}_2} |\nabla u_\varepsilon(x)| \le C_6 \max_{x \in \overline{\Omega}_3} u_\varepsilon(x) \le C_6 \max_{x \in \overline{\Omega}} \overline{u}(x). \tag{2.11}$$

From (2.11) we see that $|\nabla u_\varepsilon(x)|$ is uniformly bounded on $\overline{\Omega}_2$. It follows that $|\overline{f}_\varepsilon(x)|$ is uniformly bounded on $\overline{\Omega}_2$, and hence $\overline{f}_\varepsilon \in L^p(\Omega_2)$ for any $p > 1$. Since $-\Delta u_\varepsilon = \overline{f}_\varepsilon(x)$, $x \in \Omega_2$, it follows by Theorem 9.11 of [15] that there exists a positive constant C_7 independent of ε such that

$$||u_\varepsilon||_{W^{2,p}(\Omega_1)} \le C_7(||\overline{f}_\varepsilon||_{L^p(\Omega_2)} + ||u_\varepsilon||_{L^p(\Omega_2)}), \tag{2.12}$$

i.e., $||u_\varepsilon||_{W^{2,p}(\Omega_1)}$ is uniformly bounded. Taking $p > N$ such that $\beta < 1 - N/p$ and applying Sobolev's embedding inequality, we see that $||u_\varepsilon||_{C^{1,\beta}(\overline{\Omega}_1)}$ is uniformly bounded. Therefore $\overline{f}_\varepsilon \in C^\beta(\overline{\Omega}_1)$ and $||\overline{f}_\varepsilon||_{C^\beta(\overline{\Omega}_1)}$ is uniformly bounded. By Schauder's interior estimate theorem (see [15, Chapter 1, p. 2]), we see that there exists a positive constant C_8 independent of ε such that

$$||u_\varepsilon||_{C^{2,\beta}(\overline{\Omega'})} \le C_8(||u_\varepsilon||_{C(\overline{\Omega}_1)} + ||\overline{f}_\varepsilon||_{C^\beta(\overline{\Omega}_1)}). \tag{2.13}$$

It follows that $||u_\varepsilon||_{C^{2,\beta}(\overline{\Omega'})}$ is uniformly bounded.

From the above proof, we see that $||u_\varepsilon||_{C^{2,\beta}(\overline{\Omega'})}$ is uniformly bounded for arbitrary $\Omega' \subset\subset \Omega$. Using Ascoli–Arzela's theorem and the diagonal sequential process, we see that $\{u_\varepsilon\}$ has a subsequence that converges uniformly in the $C^2(\overline{\Omega'})$ norm to function $u \in C^2(\Omega)$ and u satisfies the equation of (2.4). From the fact $u_\varepsilon \in [\omega_\varepsilon, \overline{u}]$ and Lemma 2.7, we get $\omega(x) \le u(x) \le \overline{u}(x)$ for any $x \in \Omega$, which implies that $\lim_{x\to\partial\Omega} u(x) = 0$. Let $u|_{\partial\Omega} = 0$; thus, we get a solution $u \in C^2(\Omega)\bigcap C(\overline{\Omega})$. Applying Shauder's interior regularity theorem we see that $u \in C^{2+\beta}(\Omega)\bigcap C(\overline{\Omega})$ and thus, Lemma 2.8 is proved.

3. Proof of theorems.

3.1. Proof of Theorem 1.1. Since $p = 2$, the change of variable $\nu = e^{\lambda u} - 1$ transforms the problem (1.1) into the equivalent one

$$\begin{cases} -\Delta\nu = \dfrac{\lambda^{1+\alpha}}{\ln^{\alpha}(\nu+1)} + \lambda^{1+\alpha}h(\nu) + \lambda\sigma\nu + \lambda\sigma, \quad 0 < \nu \quad \text{in } \Omega, \\ \nu|_{\partial\Omega} = 0, \end{cases} \tag{3.1}$$

where $h(\nu) = \frac{\nu}{\ln^{\alpha}(\nu+1)}$.

If $\lambda\sigma \ge \lambda_1$, we prove that (3.1) has no solution in $H_0^1(\Omega)$, i.e., (1.1) has no solution in $H_0^1(\Omega)$. We assume the contrary; thus (3.1) has one solution in $H_0^1(\Omega)$. Since $-\Delta\nu > \lambda\sigma\nu + \lambda\sigma$ in Ω, we obtain

$$\int_\Omega (-\Delta\nu)\phi_1\,dx \ge \lambda\sigma\int_\Omega \nu\phi_1\,dx + \lambda\sigma\int_\Omega \phi_1. \tag{3.2}$$

Note that

$$\int_\Omega (-\Delta\nu)\phi_1\,dx = \lambda_1\int_\Omega \nu\phi_1\,dx.$$

From (3.2) we deduce that $\lambda\sigma < \lambda_1$. This is a contradiction. Therefore (1.1) has no solution in $H_0^1(\Omega)$ for $\lambda\sigma \ge \lambda_1$.

If $\lambda\sigma < \lambda_1$, we next prove that (3.1) has a unique classical solution. Since the function $g(s) = \frac{\lambda^{1+\alpha}}{\ln^{\alpha}(s+1)}$ satisfies the hypotheses (I_2)–(I_4), it follows from Lemma 2.7 that the singular boundary value problem

$$\begin{cases} -\Delta\nu = \dfrac{\lambda^{1+\alpha}}{\ln^{\alpha}(\nu+1)}, \quad 0 < \nu \quad \text{in } \Omega, \\ \nu|_{\partial\Omega} = 0 \end{cases} \tag{3.3}$$

has a unique solution $\omega_1 \in C(\overline{\Omega}) \bigcap C^{2+\beta}(\Omega)$ which is a subsolution of (3.1).

Now we construct a supersolution. For $0 < \alpha \leq 1$, note that

$$\lim_{s \to 0+} h(s) = \begin{cases} 1, & \alpha = 1, \\ 0, & 0 < \alpha < 1. \end{cases}$$

Then there exists a positive constant C_9 such that

$$\lambda^{1+\alpha} h(s) \leq C_9 + \frac{\lambda_1 - \lambda\sigma}{2} s \qquad \forall s > 0.$$

Then $\overline{\nu}_1 = \omega_1 + \frac{1}{2}(\lambda_1 + \lambda\sigma)\nu_1 + (C_9 + \lambda\sigma)e$ is a supersolution of (3.1) for $0 < \alpha \leq 1$, where e is the solution of (2.1) and ν_1 is the unique solution in $C^{2+\beta}(\overline{\Omega})$ of the problem (see [14, Theorem 3.2, p. 128])

$$\begin{cases} -\Delta u = \dfrac{1}{2}(\lambda_1 + \lambda\alpha)u + |\omega_1|_\infty + (C_9 + \lambda\sigma)|e|_\infty & \text{in } \Omega, \\ u|_{\partial\Omega} = 0. \end{cases} \tag{3.4}$$

From the positive lemma of Keller and Cohen [19, p. 1363], it follows that $0 < \nu_1$ in Ω.

For $\alpha > 1$, we rewrite (3.1) in the following form:

$$\begin{cases} -\Delta \nu = \dfrac{\lambda^{1+\alpha}}{\nu^\alpha} + \dfrac{\lambda^{1+\alpha}}{\ln^\alpha(\nu+1)} + \lambda^{1+\alpha}\overline{h}(\nu) + \lambda\sigma\nu + \lambda\sigma, \quad 0 < \nu \quad \text{in } \Omega, \\ \nu|_{\partial\Omega} = 0, \end{cases} \tag{3.5}$$

where $\overline{h}(\nu) = \frac{\nu}{\ln^\alpha(\nu+1)} - \frac{1}{\nu^\alpha}$.

Since $\alpha > 1$, we have $\lim_{s \to 0+} \overline{h}(s) = -\infty$. Then there exists a positive constant C_{10} such that

$$\lambda^{1+\alpha}\overline{h}(s) \leq C_{10} + \frac{\lambda_1 - \lambda\sigma}{2} s \qquad \forall s > 0. \tag{3.6}$$

Then $\overline{\nu}_2 = \omega_2 + \frac{1}{2}(\lambda_1 + \lambda\sigma)\nu_2 + (C_{10} + \lambda\sigma)e$ is a supersolution of (3.1), where ω_2 is the unique classical solution of the problem (Lemma 2.7)

$$\begin{cases} -\Delta u = \lambda^{1+\alpha}\left(\dfrac{1}{u^\alpha} + \dfrac{1}{\ln^\alpha(u+1)}\right), \quad 0 < u \quad \text{in } \Omega, \\ u|_{\partial\Omega} = 0, \end{cases} \tag{3.7}$$

and ν_2 is the unique solution in $C^{2+\beta}(\overline{\Omega})$ of the problem (see [14, 19])

$$\begin{cases} -\Delta u = \dfrac{1}{2}(\lambda_1 + \lambda\sigma)u + |\omega_2|_\infty + (C_{10} + \lambda\sigma)|e|_\infty, \quad 0 < u \quad \text{in } \Omega, \\ u|_{\partial\Omega} = 0. \end{cases} \tag{3.8}$$

Obviously, $\omega_1 \leq \overline{\nu}_1$, $\omega_1 \leq \omega_2 \leq \overline{\nu}_2$ on $\overline{\Omega}$.

Since

$$h(s) = \frac{\lambda^{1+\alpha} s}{\ln^\alpha(s+1)} \notin C^1([0,\infty),\, [0,\infty)),$$

we consider the perturbed problem of (3.1)

$$\begin{cases} -\Delta\nu = \dfrac{\lambda^{1+\alpha}}{\ln^{\alpha}(\nu+\varepsilon+1)} + \dfrac{\lambda^{1+\alpha}\nu}{\ln^{\alpha}(\nu+\varepsilon+1)} + \lambda\sigma\nu + \lambda\sigma, \quad 0 < \nu \quad \text{in } \Omega, \\ \nu|_{\partial\Omega} = 0. \end{cases} \tag{3.9}$$

By the same proof as in Lemma 2.8, we can obtain that (3.1) has a solution $\nu \in C(\overline{\Omega}) \bigcap C^{2+\beta}(\Omega)$, i.e., (1.1) has at least one solution $u = \frac{1}{\lambda}\ln(\nu+1)$.

Finally we prove that (1.1) has a unique classical solution (in $C(\overline{\Omega}) \bigcap C^{2+\beta}(\Omega)$). We assume on the contrary that (1.1) has two classical solutions u_1 and u_2, and $\{x|\, u_1(x) > u_2(x)\} \neq \emptyset$. Let $w = u_1 - u_2$, then there exists $x_0 \in \Omega$ such that $0 < w(x_0) = \max_{x\in\Omega} w(x)$. By the basic facts (see [15]) $\nabla w(x_0) = 0$, $\Delta w(x_0) \leq 0$. But this is impossible because

$$-\Delta w(x_0) = \frac{1}{u_1^{\alpha}(x_0)} - \frac{1}{u_2^{\alpha}(x_0)} < 0.$$

This contradiction implies that the problem (1.1) has a unique classical solution. The proof is complete. □

3.2. Proof of Theorem 1.2. To apply Lemma 2.8, note that the unique solution ω of the problem (2.2) is a subsolution of (1.1). For an arbitrary positive constant C, since $0 < p < 2$, by the basic facts

$$\frac{s^p}{s^2+C} \leq \frac{p^{p/2}(2-p)^{(2-p)/2}}{2C^{1-p/2}} \qquad \forall s \geq 0$$

and

$$p^{p/2}(2-p)^{(2-p)/2} < 2,$$

we have

$$s^p \leq \frac{s^2}{C^{1-p/2}} + C^{p/2} \qquad \forall s \geq 0. \tag{3.10}$$

Now we consider the problem

$$\begin{cases} -\Delta u = \dfrac{1}{u^{\alpha}} + \lambda C^{p/2-1}|\nabla u|^2 + \lambda C^{p/2} + \sigma, \quad 0 < u \quad \text{in } \Omega, \\ u|_{\partial\Omega} = 0. \end{cases} \tag{3.11}$$

Let

$$\lambda(\lambda C^{p-1} + \sigma C^{p/2-1}) < \lambda_1. \tag{3.12}$$

It follows from Theorem 1.1 that (3.11) has a unique classical solution $\overline{u}$ which is a supersolution of the problem (1.1) for $0 < p < 2$. Moreover, we can easily prove that $\omega \leq \overline{u}$ on $\overline{\Omega}$. From Lemma 2.8, we know that the problem (1.1) has at least one solution. Using the same proof as that in Theorem 1.1, we can obtain that the problem (1.1) has just one classical solution for $0 < p < 2$.

Now we analyze the inequality (3.12). For $0 < p < 1$, given every $\lambda > 0$ and every $\sigma \geq 0$, we can choose C large enough such that (3.12) holds. Thus the problem has a unique classical solution for all $\lambda \geq 0$ and $\sigma \geq 0$, i.e., $\overline{\lambda} = \infty$.

For $p = 1$, if $\lambda^2 < \lambda_1$, then given every $\sigma \geq 0$, we can choose C large enough such that (3.12) holds. For $1 < p < 2$, if $\sigma = 0$, then given every $\lambda > 0$, we can choose C satisfying $C^{p-1} < \lambda_1/\lambda^2$ such that (3.12) holds, and if $\sigma > 0$, let

$$C = \left(\frac{(2-p)\sigma}{2\lambda(p-1)}\right)^{p/2}, \tag{3.13}$$

then we can choose λ sufficiently small enough such that (3.12) holds.

To complete the proof of Theorem 1.2 for $1 \leq p < 2$ and $0 < \sigma$, let

$$A = \{\lambda > 0 | \text{ (1.1) has a unique solution } u_\lambda\}, \qquad \overline{\lambda} = \sup A.$$

By the above proof, we see $A \neq \emptyset$. It suffices to prove that if $\widetilde{\lambda} \in A$, then $(0, \widetilde{\lambda}) \subset A$. Let $u_{\widetilde{\lambda}}$ be a unique solution of (1.1) for $\lambda = \widetilde{\lambda}$. Obviously, $u_{\widetilde{\lambda}}$ is a supersolution of (1.1) for $\lambda < \widetilde{\lambda}$, and ω is a subsolution of (1.1) for $\lambda < \widetilde{\lambda}$. Moreover, we can easily obtain $\omega \leq u_{\widetilde{\lambda}}$ in Ω. By Lemma 2.8, (1.1) has at least one classical solution for $0 < \lambda < \widetilde{\lambda}$. By the proof of Theorem 1.1, (1.1) has a unique classical solution $u_{\widetilde{\lambda}}$ for $0 < \lambda < \widetilde{\lambda}$. By the definition of $\overline{\lambda}$, we know that (1.1) has no classical solution for $\lambda > \overline{\lambda}$. We assert that $\overline{\lambda} < \infty$. In fact, since $1 < p < 2$, $0 < \sigma$ and

$$\begin{cases} -\Delta u > \lambda |\nabla u|^p + \sigma & \text{in } \Omega, \\ u|_{\partial\Omega} = 0. \end{cases} \tag{3.14}$$

Using Lemma 2.4 there exists $\lambda^* < \infty$ such that problem (3.11) has no solution in $C^2(\Omega) \bigcap C(\overline{\Omega})$ for $\sigma\lambda^{\frac{1}{p-1}} > \lambda^*$. It is following that $\overline{\lambda}$ is finite for $1 < p < 2$ and $0 < \sigma$. The proof is complete. □

3.3. Proof of Theorem 1.3. Let u be a classical solution of (1.1). Obviously the unique solution ω of (2.2) is a subsolution of (1.1), $u \geq \omega$ in Ω. By Lemma 2.3, there exists a positive constant C_3 such that if $\alpha > 1$, then $C_3\phi^{2/(1+\alpha)} \leq u$ on Ω.

First we prove 1^0, i.e., if $\alpha > 1$, then u is not in $C^1(\overline{\Omega})$.

This proof is the same as the one in [7, Theorem 2]. For every $x_0 \in \partial\Omega$,

$$\frac{\partial u(x_0)}{\partial \vec{n}} = \lim_{s\to 0+} \frac{u(x_0 + s\vec{n}) - u(x_0)}{s} \leq C_3 \frac{\partial \phi_1(x_0)}{\partial \vec{n}} \left(\frac{1}{\lim_{s\to 0+} \phi_1(x_0 + s\vec{n})}\right)^{\frac{\alpha-1}{\alpha+1}} = -\infty.$$

So u is not in $C^1(\overline{\Omega})$.

Secondly, we prove 2^0, i.e., if $\alpha < 3$, then $u \in H_0^1(\Omega)$. By the proof of Lemma 2.8 and Theorems 1.1 and 1.2, it follows that $\{u_\varepsilon\}$ has a subsequence which we may assume is the sequence itself, which converges uniformly in the $C^2(\overline{\Omega}')$ norm to u for an arbitrary $C^{2+\beta}$-smooth domain $\Omega' \subset\subset \Omega$, where u_ε satisfies

$$\begin{cases} -\Delta u_\varepsilon = \dfrac{1}{(u_\varepsilon + \varepsilon)^\alpha} + \lambda |\nabla u_\varepsilon|^p + \sigma, \quad 0 < u_\varepsilon \quad \text{in } \Omega, \\ u_\varepsilon|_{\partial\Omega} = 0 \end{cases} \tag{3.15}$$

and $u_\varepsilon \in C^{2+\beta}(\overline{\Omega})$. Moreover, it follows by Lemmas 2.3 and 2.7 that $\omega_\varepsilon \leq u_\varepsilon \leq u$ on $\overline{\Omega}$, where ω_ε is the unique solution of the problem

$$\begin{cases} -\Delta u = \dfrac{1}{(u + \varepsilon)^\alpha}, \quad 0 < u \quad \text{in } \Omega, \\ u|_{\partial\Omega} = 0. \end{cases} \tag{3.16}$$

Since

$$\int_\Omega |\nabla u_\varepsilon|^2\,dx \le \int_\Omega \frac{u_\varepsilon}{(u_\varepsilon+\varepsilon)^\alpha}\,dx + \lambda\int_\Omega u_\varepsilon|\nabla u_\varepsilon|^p\,dx + \sigma|\Omega||u|\infty, \tag{3.17}$$

and $u \in C(\overline{\Omega})$, $u|_{\partial\Omega} = 0$, Ω is a bounded domain with $C^{2+\beta}$ boundary, and given $\lambda > 0$ and $\sigma \ge 0$, we know that there exists a neighborhood U of $\partial\Omega$ such that if $\Sigma = U \bigcap \Omega$, then

$$\lambda u_\varepsilon \le \lambda u \le \frac{1}{2} \quad \text{in } \Sigma. \tag{3.18}$$

Furthermore, by Lemma 2.7 we have

$$\omega_\varepsilon + \varepsilon \ge \omega \quad \text{on } \overline{\Omega}. \tag{3.19}$$

Consequently,

$$\int_\Omega \frac{u_\varepsilon}{(u_\varepsilon+\varepsilon)^\alpha}\,dx < \int_\Omega u_\varepsilon^{1-\alpha}\,dx < \int_\Omega u^{1-\alpha}\,dx \quad \text{for } \alpha \in (0,1], \tag{3.20}$$

and for $\alpha \in (1,3)$,

$$\int_\Omega \frac{u_\varepsilon}{(u_\varepsilon+\varepsilon)^\alpha}\,dx < \int_\Omega \frac{1}{(u_\varepsilon+\varepsilon)^{\alpha-1}}\,dx < \int_\Omega \frac{1}{(\omega_\varepsilon+\varepsilon)^{\alpha-1}}\,dx < \int_\Omega \frac{1}{\omega^{\alpha-1}}\,dx. \tag{3.21}$$

So, by Lemmas 2.3 and 2.2, we have

$$\int_\Omega \frac{1}{\omega^{\alpha-1}}\,dx \le \frac{1}{C_3^{\alpha-1}}\int_\Omega \left(\frac{1}{\phi_1}\right)^{\frac{2(\alpha-1)}{1+\alpha}}\,dx < \infty. \tag{3.22}$$

It follows by (3.17)–(3.22) that

$$\int_\Omega |\nabla u_\varepsilon|^2\,dx \text{ is bounded independently of } \varepsilon. \tag{3.23}$$

Thus $u \in H_0^1(\Omega)$.

Finally we prove 3^0. We assume on the contrary that $\alpha \ge 3$, $u \in H_0^1(\Omega)$. It follows by Green's identity that

$$\int_\Omega |\nabla u|^2\,dx = \int_\Omega \frac{1}{u^{\alpha-1}}\,dx + \lambda\int_\Omega u|\nabla u|^p\,dx + \sigma\int_\Omega u\,dx. \tag{3.24}$$

Consequently

$$\int_\Omega \frac{1}{u^{\alpha-1}}\,dx \text{ is bounded.} \tag{3.25}$$

But this is impossible, because under the assumptive conditions of 3^0 we will obtain (in the following proof)

$$u \le C_{11}\phi_1^{2/(1+\alpha)} \quad \text{on } \overline{\Sigma}, \tag{3.26}$$

where $\Sigma \subset \Omega$ is a neighborhood of $\partial\Omega$ and C_{11} is a positive constant and $\alpha > 1$. By the proof of Lemma 2.2 in [7] and (3.26), we obtain

$$\int_\Omega \frac{1}{u^{\alpha-1}}\,dx = \infty \quad \text{for } \alpha \ge 3. \tag{3.27}$$

This contradiction implies that 3^0 is true.

In the remainder of this paper, we prove the inequality (3.26). Let us return to the proof of Theorems 1.1 and 1.2. We have known that (1.1) has just one classical solution for all $\lambda \geq 0$ when the condition (c_2) or (c_3) is satisfied, and the proof of Theorem 1.2 is the same as that of Theorem 1.1 because of (3.10)–(3.12). Thus it suffices to prove (3.26) when $p = 2$ in (1.1) and (3.1). Using $p = 2$ and (3.5)–(3.9), we have known that $\overline{\nu}_2 = \omega_2 + (\lambda\sigma + C_6)e + \frac{1}{2}(\lambda_1 + \lambda\sigma)\nu_2$ is a supersolution of (3.1) for $\alpha > 1$, and

$$u \leq \frac{1}{\lambda}\ln(\overline{\nu}_2 + 1) \leq \frac{1}{\lambda}\overline{\nu}_2 \quad \text{on } \overline{\Omega}. \tag{3.28}$$

Since ν_2 is the unique solution of the problem (3.8), using Hopf's maximum principle (see [15]), we have

$$\frac{\partial \nu_2(x)}{\partial \vec{n}} < 0 \quad \forall\, x \in \partial\Omega. \tag{3.29}$$

Thus we obtain that there exist two positive constants C_{12} and C_{13} such that

$$C_{12}\phi_1 \leq \nu_2 \leq C_{13}\phi_1 \quad \text{on } \overline{\Omega}. \tag{3.30}$$

Now we estimate ω_2. Since ω_2 is the unique solution of the problem (3.7), let

$$\overline{\omega}_2 = M\phi_1^{2/(1+\alpha)},$$

where M is a large positive constant to be chosen. Then we have

$$-\Delta\overline{\omega}_2 = \frac{2M(\alpha-1)|\nabla\phi_1|^2}{(1+\alpha)^2\phi_1^{2\alpha/(1+\alpha)}} + \frac{2\lambda_1 M}{1+\alpha}\phi_1^{2/(1+\alpha)} \quad \text{on } \Omega.$$

We need for the following inequality

$$-\Delta\overline{\omega}_2 \geq \frac{\lambda^{1+\alpha}}{\overline{\omega}_2^{\alpha}} + \frac{\lambda^{1+\alpha}}{\ln^{\alpha}(1+\overline{\omega}_2)} \quad \text{in } \Omega. \tag{3.31}$$

That is

$$M^{1+\alpha}\left(\frac{2(\alpha-1)|\nabla\phi_1|^2}{(1+\alpha)^2} + \frac{2\lambda_1}{1+\alpha}\phi_1^2\right) \geq \lambda^{1+\alpha} + \lambda^{1+\alpha}\left(\frac{M\phi_1^{2/(1+\alpha)}}{\ln(1+M\phi_1^{2/(1+\alpha)})}\right)^{\alpha} \quad \text{in } \Omega. \tag{3.32}$$

Since $\alpha > 1$, $\nabla\phi_1(x) \neq 0$ for all $x \in \partial\Omega$, and the function $h(s) = \frac{s^{\alpha}}{\ln^{\alpha}(1+s)}$ has the properties that $h \in C^1(0,\infty)$ is strictly increasing and $\lim_{s\to 0+} h(s) = 1$. Thus we have

$$\min_{x\in\overline{\Omega}}\left(\frac{2(\alpha-1)}{(1+\alpha)^2}|\nabla\phi_1(x)|^2 + \frac{2\lambda_1}{1+\alpha}\phi_1^2(x)\right) = C_0 > 0,$$

and there exists $C_{14} > 0$ such that

$$h(s) \leq C_{14} + \frac{C_0 s^{\alpha}}{2|\phi_1|_{\infty}^{2/(1+\alpha)}} \quad \forall\, s > 0. \tag{3.33}$$

It follows that

$$h(M\phi_1^{2/(1+\alpha)}) \le h(M|\phi_1|_\infty^{2/(1+\alpha)}) \le C_{14} + \frac{1}{2}C_0 M^\alpha \quad \text{on } \overline{\Omega}.$$

Thus, there exists a M_0 such that for all $M \ge M_0$, $\overline{\omega}_2 = M\phi_1^{2/(1+\alpha)}$ satisfying (3.32), i.e., $\overline{\omega}_2$ is a supersolution of the problem (3.7). We can obtain

$$\omega_2 \le \overline{\omega}_2 = M_0\phi_1^{2/(1+\alpha)} \quad \text{on } \overline{\Omega}. \tag{3.34}$$

By Lemma 2.1 and (3.34), (3.30), we know that there exists a positive constant C_{15} such that

$$\overline{\nu}_2 \le C_{15}(\phi_1^{2/(1+\alpha)} + \phi_1) \quad \text{on } \Omega. \tag{3.35}$$

Since $2/(1+\alpha) < 1$ and $s \le s^{2/(1+\alpha)}$ for all $s \in (0,1)$, then we obtain (3.26). The proof of Theorem 1.3 is complete. □

Acknowledgment. The authors are indebted to the references for the valuable results.

REFERENCES

[1] W. Fulks and J. S. Maybee, *A singular nonlinear equation*, Osaka J. Math., 12 (1960), pp. 1–19.
[2] C. A. Stuart, *Existence and approximation of solution of nonlinear elliptic equations*, Math. Z., 147 (1976), pp. 53–63.
[3] M. G. Crandall, P. H. Rabinowitz, and L. Tartar, *On a Dirichlet problem with a singular nonlinearity*, Comm. Partial Differential Equations, 2 (1977), pp. 193–222.
[4] S. M. Gomes, *On a singular nonlinear elliptic problem*, SIAM J. Math. Anal., 17 (1986), pp. 1359–1369.
[5] M. M. Coclite and G. Palmieri, *On a singular nonlinear dirichlet problem*, Comm. Partial Differential Equations, 14 (1989), pp. 1315–1327.
[6] H. Usami, *On a singular elliptic boundary value problem in a ball*, Nonlinear Anal., 13 (1989), pp. 1163–1170.
[7] A. C. Lazer And P. J. McKenna, *On a singular elliptic boundary value problem*, Proc. Amer. Math. Soc., 111 (1991), pp. 721–730.
[8] S. Cui, *Positive solutions for Dirichlet problems associated to semilinear elliptic equations with singular nonlinearity*, Nonlinear Anal., 21 (1993), pp. 181–190.
[9] A. C. Lazer And P. J. McKenna, *On a singular boundary value problem for the Monge-Ampere operator*, J. Math. Anal. Appl., 197 (1996) pp. 341–362.
[10] M. M. Coclite, *On a singular nonlinear Dirichlet problem* III, Nonlinear Anal., 21 (1993), pp. 547–564.
[11] J. I. Diaz, J. M. Morel, and L. Oswald, *An elliptic equation with singular nonlinearity*, Comm. Partial Differential Equations, 12 (1987), pp. 1333–1344.
[12] M. M. Coclite, *On a singular nonlinear Dirichlet problem* II, Boll. Un. Math. Ital., B (7), 5 (1991), pp. 955–975.
[13] H. Amann, *Existence and multiplicity theorems for semilinear elliptic boundary value problems*, Math. Z., 150 (1976), pp. 567–597.
[14] O. A. Ladyzenskaja and N. N. Ural'ceva, *Equations aux Derivees Partielles de Type Elliptique*, Dunod, Paris, 1968 (English translation).
[15] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Springer, Berlin, 1983.
[16] A. Nachman and A. Callegari, *A nonlinear singular boundary value problem in the theory of pseudoplastic fluids*, SIAM J. Appl. Math., 38 (1980), pp. 275–281.
[17] N. E. Alaa and M. Pierre, *Weak solutions of some quasilinear elliptic equations with data measures*, SIAM J. Math. Anal., 24 (1993), pp. 23–35.
[18] P.-L. Lions, *On the existence of positive solutions of semilinear elliptic equations*, SIAM Rev., 24 (1982), pp. 441–467.

[19] H. B. Keller and D. S. Cohen, *Some positive problems suggested by nonlinear heat generation*, J. Math. Mech., 16 (1967), pp. 1361–1376.
[20] Z. Zhang, *On a Dirichlet problem with a singular nonlinearity*, J. Math. Anal. Appl., 194 (1995), pp. 103–113.
[21] Z. Zhang, *Nonexistence of positive classical solutions of a singular nonlinear Dirichlet problem with a convection term*, Nonlinear Anal., 27 (1996), pp. 957–961.
[22] Z. Zhang, *On singular Dirichlet problems for semilinear elliptic equation of second order*, Northeast. Math. J., 11 (1995), pp. 31–38.
[23] Z. Zhang, *Nonlinear elliptic equations with singular boundary conditions*, J. Math. Anal. Appl., 216 (1997), pp. 390–397.

SIAM J. MATH. ANAL.
Vol. 32, No. 5, pp. 929–962

ALTERNATE EVANS FUNCTIONS AND VISCOUS SHOCK WAVES*

SYLVIE BENZONI-GAVAGE†, DENIS SERRE‡, AND KEVIN ZUMBRUN§

Abstract. The Evans function is known as a helpful tool for locating the spectrum of some variational differential operators. This is of special interest regarding the stability analysis of traveling waves, such as reaction-diffusion waves, solitary waves, viscous shock waves, etc., and has been used in numerous contexts. The first aim of this paper is to present an overview of the various ways to define an Evans function for an abstract differential operator. Not all of these alternatives are new, but we show consistent connections between them. Subsequently, we focus on viscous shock waves, extending the work of Gardner and Zumbrun in several directions. In particular, we (i) show some advantages of alternate Evans functions in practical computations, (ii) carry out a refined analysis in case of neutral stability, and (iii) show how to treat systems of size $n > 2$, thus resolving a problem left open by Gardner and Zumbrun.

Key words. traveling waves, asymptotic stability, viscous conservation laws

AMS subject classifications. 34L15, 35K45, 35L67

PII. S0036141099361834

1. Introduction.

The Evans function is a tool that extends the notion of characteristic polynomial to infinite-dimensional operators, such as variational differential operators. It was first introduced in a special case by Evans [5, 6, 7, 8]. In its present generality it is due to [1]. Given an operator L, an Evans function D is a function of the frequency λ, which is analytic away from the essential spectrum of L and vanishes only on the point spectrum of L. Though not explicitly evaluable in any but very simple cases, it can nonetheless yield a great deal of information through various topological considerations. It has been successfully applied to the analysis of stability of traveling waves; see, e.g., [14, 1, 20, 2, 15, 12, 17], etc.

In the first part of this paper, we propose various ways to define an Evans function in a rather abstract framework. A special effort is made to clarify the relations between these alternatives. In view of the application to viscous shock waves, we consider a second order differential operator L (attaining exponentially fast some limits $L_\pm$ at $\pm\infty$), but the basic principles can be applied to other kinds of operators. The essential spectrum of L can be localized through standard arguments. In most generality, it should lie to the left of the spectra of $L_\pm$, which can be determined by Fourier transform. An Evans function is aimed at localizing eigenvalues of L away from its essential spectrum, in particular in the right half-plane. The starting point of the construction is well known. It consists in formulating the eigenvalue equations $Lw = \lambda w$ as a (variational) dynamical system, depending on the parameter λ, $W' = \mathbb{A}(x;\lambda)\, W$. In our case, the eigenfunctions w are searched in $H^2(\mathbb{R};\mathbb{C}^n)$, W is valued in $\mathbb{C}^{2n}$, and $\mathbb{A}(x;\lambda)$ is a square matrix of size $2n$. The basic assumption is that the dimensions of the stable and unstable manifolds of the dynamical system are complementary (for a certain range of the parameter λ). In particular, this requires

*Received by the editors September 24, 1999; accepted for publication (in revised form) August 22, 2000; published electronically January 5, 2001.
http://www.siam.org/journals/sima/32-5/36183.html
†CNRS and ENS Lyon, UMPA, 46, allée d'Italie, F-69364 Lyon Cedex 07, France (benzoni@umpa.ens-lyon.fr).
‡ENS Lyon, UMPA, 46, allée d'Italie, F-69364 Lyon Cedex 07, France (serre@umpa.ens-lyon.fr).
§Department of Mathematics, Indiana University, Rawles Hall, Bloomington, IN 47405 (kzumbrun@indiana.edu).

that the limit matrices $\mathbb{A}_\pm(\lambda)$ do not have a null space. This is true away from the essential spectrum of L. Then by choosing a parametrization of these manifolds, we see that they intersect in a nontrivial way if and only if a certain Wronskian $D(\lambda)$ vanishes. This is the classical way to define an Evans function. It is clear that $\lambda \notin \sigma_{\mathrm{ess}}(L)$ is an eigenvalue of L if and only if $D(\lambda) = 0$. Further refinements show that the order of vanishing of D equals the multiplicity of the eigenvalue (see [1]). An alternative way of defining an Evans function is rather to consider the *adjoint operator* L^*, using the fact that L and L^*, which are real-valued, have the same eigenvalues. This idea yields a "dual" Evans function, which we shall denote by D_*. It will appear to be of simpler use than the classical one, at least for viscous conservation laws. Another approach is of "mixed" type. It consists in characterizing the stable (or the unstable) manifold of $W' = \mathbb{A}(x;\lambda)\, W$ as the orthogonal of the stable (or the unstable) manifold of the *adjoint dynamical system* $Z' = -\mathbb{A}(x;\lambda)^*\, Z$. By choosing a parametrization of these manifolds, we obtain two alternate Evans functions, defined as "Gramian determinants" (instead of a Wronskian), which we shall denote by $D_\pm$ (the subscripts $\pm$ here do not refer to any limit at $\pm\infty$ but only to the way the basis of the manifolds are numbered). By convention of our notations, the size of D_- (respectively, D_+) is the dimension of the stable (respectively, unstable) manifold. It is well suited when this number is small. In our case, this number is equal to n. We shall nevertheless show that the use of such D_- of D_+ simplifies the calculation in some respects. It also makes possible the extension to situations in which the standard construction is not possible, for example, as in the case when the dynamical system is infinite-dimensional. (See the treatment of semidiscrete difference schemes in [3].) It is to be noted that this kind of Evans function generalizes the one used in [20]. It was first introduced by Swinton [23], and was used (in a different) form in [2]. The construction of Swinton, which holds away from branch points of the asymptotic matrices $\mathbb{A}_\pm(\lambda)$, has been related to the classical one in [4]. We attempt here to give a unified presentation, which does not preclude branch points and can be useful in various contexts.

The remaining part of the paper is devoted to viscous shock waves, but some general ideas can be translated to other contexts. Motivating problems concerning viscous shock waves come from gas dynamics, MHD, and also from the numerics of hyperbolic systems of conservation laws. The application of an Evans function method to viscous shock waves was first carried out in [12]. Gardner and Zumbrun used it to determine necessary conditions for stability in 2×2 systems. An important aspect of this analysis, which we shall not discuss much here, was the extension of the analytic framework of [1] to situations, such as occurs for viscous shock waves, in which the essential spectrum of the operator L accumulates at the imaginary axis and hence there is no spectral gap between essential spectrum and the unstable half-plane $\{\mathrm{Re}\,\lambda > 0\}$. This necessitates a refined analysis justifying analytic extension to the boundary, specifically, the "gap lemma" established independently in [12, 17]. A second aspect of the analysis in [12] was the actual calculation of stability conditions, following the model of [14, 20]; it is primarily this aspect that will concern us here. The basic approach is a natural one: observing that $D(0) = 0$, corresponding to a translational eigenvalue at $\lambda = 0$, one determines the sign of $D'(0)$ by appropriate low-frequency analysis. One then determines the sign of $D(\cdot)$ as $\lambda \to +\infty$ along the real axis (note: $D(\lambda)$ can be chosen to be real for real λ by symmetry) by separate, high-frequency analysis and compares the two. This gives a parity for the number of unstable zeros of D/eigenvalues of L; if the parity is odd, one can conclude instability.

We extend the approach of [12] in several directions.[1] First, we carry out the above calculation for the alternate Evans functions D_* and D_+. We point out that these rather specialized computations can be helpful to a reader working in a different field. In particular, they have been helpful with the work of the first author concerning semidiscrete profiles [3]. Along the way, we give a simplified treatment of the high-frequency limit, based on *homotopy* rather than rescaling/invariant sets as in [1, 12]. Next, we consider the case of neutral instability $D'(0) = 0$. We give a formula for $D''(0)$, which has similarities with Kapitula's formulas in [16]. This yields interesting examples of instability despite even parity of unstable roots. Finally, we prove a key linear algebraic result conjectured in [12], allowing the extension from 2×2 to $n \times n$ systems, also carrying out one or two calculations in this general setting (notably, the under-compressive case).

The paper is organized as follows. In section 2, we recall the definition of the standard Evans function D and we introduce the alternate ones, namely, D_* and $D_\pm$. In section 3, we recall the material necessary to the stability analysis of viscous shocks waves by means of these Evans functions. In section 4, we apply the dual Evans function D_* to viscous shock waves. First we present the computation of $D'_*(0)$, which differs significantly from the one of $D'(0)$. Then we use a homotopy method, instead of the rescaling method in [12], to derive the sign of $D_*(\lambda)$ for large real λ. In section 5, we consider the mixed Evans function D_+. We also compute the sign of $D_+(+\infty)$ by a means of a homotopy method. The main novelty lies in the computation of $D'_+(0)$, which uses the relation between the adjoint ODE and the adjoint PDE. This emphasizes the importance of this relation, which has already been used for different purposes in the literature (see [16, 24]). In section 6, we study the transition occurring when $D'(0) = 0$. In section 7, we complete the work of Gardner and Zumbrun [12] for $n \times n$ systems. This extension is important in view of the applications.

2. Definition of various Evans functions.

2.1. The classical Evans function. We consider a second order differential operator

$$Lw = Bw'' + Cw' + Dw,$$

acting on vector fields $w : \mathbb{R} \to \mathbb{R}^n$. The real-valued $n \times n$ matrices B, C, D are assumed to depend smoothly on x. Moreover, they are supposed to have limits $B_\pm$, $C_\pm$, $D_\pm$, at $\pm\infty$, with exponential rate of decay $\alpha > 0$, that is,

$$\begin{aligned} \|B(x) - B_+\| + \|C(x) - C_+\| + \|D(x) - D_+\| &\le k\, e^{-\alpha x}, \quad x > 0, \\ \|B(x) - B_-\| + \|C(x) - C_-\| + \|D(x) - D_-\| &\le k\, e^{\alpha x}, \quad x < 0. \end{aligned}$$

We denote by $L_\pm$ the limit operators

$$L_\pm := B_\pm \frac{d^2}{dx^2} + C_\pm \frac{d}{dx} + D_\pm.$$

We assume that L is uniformly elliptic, that is, there exists a positive number β such that

$$(B(x)v \,|\, v) \ge \beta\, |v|^2 \quad \forall (x, v) \in \mathbb{R} \times \mathbb{R}^n. \tag{2.1}$$

[1] For further extensions, in the multidimensional case, see [25].

We point out that this assumption forces the eigenvalues of $B(x)$ to be of positive real part.

The essential spectrum of L then lies in some left half-plane $\{\operatorname{Re}\lambda < \rho\}$. As a matter of fact, this holds true for $L_\pm$, as we can easily see by Fourier transform, using the estimate (2.1) at $x = \pm\infty$. Furthermore, we can use (2.1) in a rough energy estimate to show that, for $\operatorname{Re}\lambda$ large enough, λ cannot be an eigenvalue of L. Therefore, by a classical argument due to Henry [13], we find that $\sigma_{ess}(L)$ lies on the left of $\sigma(L_+) \cup \sigma(L_-)$, and thus it lies in a left half-plane. In most applications, $\sigma_{ess}(L)$ lies in $\{\operatorname{Re}\lambda < 0\}$, apart from the origin. In this case, stability properties are linked to possible eigenvalues of L of positive real part.

Since $B_\pm$ have eigenvalues of positive real part we find that, for $\operatorname{Re}\lambda \gg 1$, the solutions of $(L-\lambda)w = 0$ decaying to zero at $+\infty$ span an n-dimensional vector space, $\mathcal{E}_+(\lambda)$, as well as the solutions decaying to zero at $-\infty$, $\mathcal{E}_-(\lambda)$. This is still true by a continuity argument for $\operatorname{Re}\lambda > \rho$. Then the Evans function $\lambda \mapsto D(\lambda)$ is defined for $\operatorname{Re}\lambda > \rho$ by two (locally) analytic maps $\lambda \mapsto \mathcal{B}_\pm(\lambda)$, where $\mathcal{B}_\pm(\lambda)$ are bases of $\mathcal{E}_\pm(\lambda)$. Let us denote by $\phi_{j\pm}(\lambda)$ the elements of the bases. These are vector fields, belonging to $\ker(L-\lambda)$. The definition of the Evans function is

$$D(\lambda) := \left| \begin{array}{cccccc} \phi_{1-} & \cdots & \phi_{n-} & \phi_{1+} & \cdots & \phi_{n+} \\ \phi'_{1-} & \cdots & \phi'_{n-} & \phi'_{1+} & \cdots & \phi'_{n+} \end{array} \right|_{x=0}, \tag{2.2}$$

where the primes denote x-derivatives. The function D is locally analytic and vanishes at λ if and only if λ is an ordinary eigenvalue of L, that is, there exists a nontrivial $w \in L^2(\mathbb{R})$ such that $Lw = \lambda w$. Actually, by working in the exterior product $\Lambda^n(\mathbb{C}^{2n})$, it is possible to construct a function that is globally analytic and vanishes at the same points; see [1, 12]. Furthermore, it vanishes at λ with order m if and only if the generalized eigenspace (that is, the union of kernels of $(L-\lambda)^k$ in $L^2(\mathbb{R})$) is of dimension m [11, 10, 24]. Additionally, the spaces $\mathcal{E}_\pm(\lambda)$ are real when λ is real, in the sense that they are stable by complex conjugation. It turns out that their bases may be chosen in such a way that $D(\lambda)$ is real for real λ; see, e.g., [12] for further details.

This definition extends in a straightforward way to the connected component Ω of the point $\rho+1$ (i.e., the rightmost component) in the complement of the essential spectrum, where $L-\lambda : H^2 \to L^2$ is a Fredholm operator with index zero.

It is often useful to extend D to a neighborhood of Ω, especially in the case when $\rho = 0$. Such an extension, when available, does not obey the above definition, but relies instead on the "gap lemma" of Gardner and Zumbrun [12], which was proved simultaneously by Kapitula and Sandstede [17]. To discuss this point, we first rewrite the equation $Lw = \lambda w$ as a first order ODE,

$$W' = \mathbb{A}(x;\lambda)W. \tag{2.3}$$

Then $\mathbb{A}(x;\lambda)$ admits limits as $x \to \pm\infty$, denoted by $\mathbb{A}_\pm(\lambda)$. When $\lambda \in \Omega$, these are hyperbolic matrices, with n-dimensional stable/unstable subspaces denoted by $S_\pm(\lambda)$, $U_\pm(\lambda)$. Introducing

$$\Phi_{j\pm} := \begin{pmatrix} \phi_{j\pm} \\ \phi'_{j\pm} \end{pmatrix}, \quad j = 1, \dots, n,$$

these are solutions of (2.3) such that

$$U_-(\lambda) = \lim_{x\to-\infty} \operatorname{Span}\{\Phi_{1-}(x;\lambda), \dots, \Phi_{n-}(x;\lambda)\}, \tag{2.4}$$

$$S_+(\lambda) = \lim_{x\to+\infty} \operatorname{Span}\{\Phi_{1+}(x;\lambda), \dots, \Phi_{n+}(x;\lambda)\}. \tag{2.5}$$

The gap lemma shows that, whenever the spaces $S_\pm(\lambda)$ and $U_\pm(\lambda)$ have analytic extensions in some neighborhoods $\mathcal{N}_\pm$ of the origin with

$$\mathbb{C}^{2n} = S_\pm(\lambda) \oplus U_\pm(\lambda)$$

(this assumption is called "geometric separation"), the spaces $\mathcal{E}_\pm(\lambda)$ also extend analytically to a neighborhood $\mathcal{N} \subset \mathcal{N}_\pm$, the size of which depends on the decay rate $\alpha > 0$. This allows us to define $D(\lambda)$ by the above determinant (2.2) in $\mathcal{N}$. However, zeros of D in this extra domain may not be eigenvalues of L in the usual sense, since the fields $\phi_{j\pm}$ need not vanish at $\pm\infty$. They are sometimes called "effective eigenvalues" [24].

Once D is properly defined on $[0, +\infty)$, we may apply the procedure described in the introduction to find a parity for the number of unstable eigenvalues of L. It relies on the computation of $D'(0)$ (in case of translational invariance, that is, $D(0) = 0$) and of the limit $D(+\infty)$. These computations were carried out in [12], in the case when L comes from the linearization about a shock wave of a 2×2 system of viscous conservation laws. They are completed for larger systems in section 7 of this paper. We are now going to propose alternative Evans functions, which may simplify some of these computations, as we shall see in the examples of sections 4 and 5.

2.2. The dual Evans function. We consider the adjoint operator L^*, acting on co-vector fields and defined by

$$L^* z = (zB)'' - (zC)' + zD.$$

We shall assume the additional decay estimates

$$\|B''(x)\| + \|C'(x)\| \le k\, e^{-\alpha|x|}, \quad x \in \mathbb{R}.$$

We now use the well-known fact that complex conjugate of eigenvalues of L in Ω are eigenvalues of the adjoint operator L^*. However, since L has real coefficients, its spectrum is invariant by complex conjugation. Therefore, L and L^* share the same eigenvalues in Ω, and the stability properties of L may be studied through the Evans function D_* of L^*, instead of D. These both vanish at the same points, with the same order.

We can especially take advantage of using D_* instead of D when L is *conservative*, that is, when it reads

$$Lw := (Bw' - Aw)'. \tag{2.6}$$

(Here the notations differ slightly from above.) Then the adjoint operator reads

$$L^* z := (z'B)' + z'A. \tag{2.7}$$

Under natural assumptions that we shall specify later, the dual Evans function D_* can be analytically defined in a neighborhood of the origin by

$$D_*(\lambda) := \begin{vmatrix} \chi_{1-}(\lambda) & \chi'_{1-}(\lambda) \\ \vdots & \vdots \\ \chi_{n-}(\lambda) & \chi'_{n-}(\lambda) \\ \chi_{1+}(\lambda) & \chi'_{1+}(\lambda) \\ \vdots & \vdots \\ \chi_{n+}(\lambda) & \chi'_{n+}(\lambda) \end{vmatrix}_{x=0}. \tag{2.8}$$

In particular at $\lambda = 0$, the basis functions $\chi_{j\pm}$ are bounded solutions of $L^*\chi = 0$. The special form (2.7) of L^* shows that some of these functions are just constants!

The assumptions that we need are the following. Again the matrices B, A and their derivatives are assumed to converge exponentially fast at $\pm\infty$. By Fourier transform we find that $\sigma(L_\pm)$ is the set of eigenvalues of all the matrices $-\xi^2 B_\pm - i\xi A_\pm$ as $\xi \in \mathbb{R}$. We assume that the matrices

$$\mathbb{A}_\pm(\lambda) := \begin{pmatrix} 0_n & I_n \\ \lambda B_\pm^{-1} & B_\pm^{-1} A_\pm \end{pmatrix} \tag{2.9}$$

do not admit pure imaginary eigenvalues for $\operatorname{Re}\lambda > 0$. Then $\sigma_{ess}(L)$ lies within the half-space $\operatorname{Re}\lambda \le 0$. This is nothing but requiring the L^2-stability of the operator $\partial_t - L$. A local analysis at the origin shows that this implies that the eigenvalues of $A_\pm$ are real. Assuming that these eigenvalues are simple and nonzero, one obtains the geometric separation (see [12]), which in turn allows us to extend either D or D_* analytically in the vicinity of the origin. At $\lambda = 0$, the basis functions $\phi_{j\pm}$ are bounded solutions of $L\phi = 0$, that is, $B\phi' = A\phi +$ constant. They are not as simple as the (constant) $\chi_{j\pm}$.

Remark 1. The dual function D_* can be related to the classical one, D, through the following procedure. Let us denote by $\mathcal{E}(\lambda)$ and $\mathcal{F}(\lambda)$ the $2n$-dimensional spaces of solutions to $(L-\lambda)v = 0$ and $(L^*-\lambda)z = 0$, respectively. Then for all $v \in \mathcal{E}(\lambda)$ and $z \in \mathcal{F}(\lambda)$, the quantity $zBv' - z'Bv - zAv$ does not depend on x. (We use this fact elsewhere to treat mixed Evans functions; see Lemma 5.1 below.) Hence we have a bilinear form

$$\begin{aligned} \mathcal{E}(\lambda) \times \mathcal{F}(\lambda) &\to \mathbb{C}, \\ (v, z) &\mapsto \langle v, z\rangle := zBv' - z'Bv - zAv, \end{aligned}$$

which is nondegenerate. The perp map with respect to $\langle\cdot,\cdot\rangle$ thus transforms n-dimensional subspaces of $\mathcal{E}(\lambda)$ into n-dimensional subspaces of $\mathcal{F}(\lambda)$. In particular, $\mathcal{E}_\pm(\lambda)$ are transformed into $\mathcal{F}_\pm(\lambda)$, the subspaces of solutions to $(L^*-\lambda)z = 0$ decaying at $\pm\infty$.

2.3. "Mixed" Evans functions. We here consider the *adjoint ODE*

$$Z' = -\mathbb{A}(x;\lambda)^* \cdot Z, \tag{2.10}$$

where $\mathbb{A}(x;\lambda)^*$ denotes the adjoint operator of $\mathbb{A}(x;\lambda)$, thus acting on co-vectors. By definition we have

$$\mathbb{A}(x;\lambda)^* \cdot Z = Z\,\mathbb{A}(x;\bar\lambda).$$

For $\lambda \in \Omega$, the adjoint matrices $\mathbb{A}_\pm(\lambda)^*$ also have n-dimensional stable/unstable subspaces, which we denote by $S_\pm^*(\lambda)$ and $U_\pm^*(\lambda)$, and we have

$$S_+(\lambda) = U_+^*(\lambda)^\perp, \tag{2.11}$$

$$U_-(\lambda) = S_-^*(\lambda)^\perp. \tag{2.12}$$

We may choose decaying solutions to (2.10), $(\Psi_{j\pm})_{1\le j\le n}$, such that

$$U_+^*(\lambda) = \lim_{x\to+\infty} \operatorname{Span}\{\Psi_{1+}^T(x;\lambda), \dots, \Psi_{n+}^T(x;\lambda)\}, \tag{2.13}$$

$$S_-^*(\lambda) = \lim_{x\to-\infty} \operatorname{Span}\{\Psi_{1-}^T(x;\lambda), \dots, \Psi_{n-}^T(x;\lambda)\}. \tag{2.14}$$

By (2.3) and (2.10), we have that the dot products $(\Psi_i(x;\lambda)\cdot\Phi_j(x;\lambda))$ are *independent of* $\boldsymbol{x}$. Thus we see from (2.5), (2.11), and (2.13) that

$$\text{Span}\{\Phi_{1+}(x;\lambda),\dots,\Phi_{n+}(x;\lambda)\} = \text{Span}\{\Psi_{1+}(x;\lambda),\dots,\Psi_{n+}(x;\lambda)\}^{\perp} \tag{2.15}$$

and from (2.4), (2.12), and (2.14) that

$$\text{Span}\{\Phi_{1-}(x;\lambda),\dots,\Phi_{n-}(x;\lambda)\} = \text{Span}\{\Psi_{1-}(x;\lambda),\dots,\Psi_{n-}(x;\lambda)\}^{\perp}. \tag{2.16}$$

Therefore, we may define alternate Evans functions in Ω by

$$D_+(\lambda) := \det(\Psi_{i+}\cdot\Phi_{j-})_{1\le i,j\le n}, \tag{2.17}$$

$$D_-(\lambda) := \det(\Psi_{i-}\cdot\Phi_{j+})_{1\le i,j\le n}. \tag{2.18}$$

In particular D_- is nothing but the generalization of the Evans function considered by Pego and Weinstein [20] in the case when the stable manifold is 1-dimensional. It is less obvious here that $D_\pm$ are globally analytic, since individual eigenvectors are not smoothly defined in case of crossing eigenvalues. However, the complete stable/unstable manifolds, which may be represented by means of wedge products, vary analytically even where individual eigenvalues may coincide. More precisely, from the standard theory on the Evans function [1, 12], we may choose $\Phi_{j\pm}$ and $\Psi_{i\pm}$, satisfying (2.4), (2.5) and (2.13), (2.14), respectively, such that the wedge products

$$\Phi_{1\pm}\wedge\cdots\wedge\Phi_{n\pm},\quad \Psi_{1\pm}\wedge\cdots\wedge\Psi_{n\pm}$$

define analytic functions from $\Omega\cup\mathcal{N}$ to $P(\Lambda^n(\mathbb{C}^{2n}))$, the manifold of projectivized n-powers of $\mathbb{C}^{2n}$. Therefore, the analytic dependence of $D_\pm$ (up to a scalar nonvanishing function) will follow from the formula (see also Proposition 2.2 below)

$$D_\pm(\lambda) = (\Psi_{1\pm}\wedge\cdots\wedge\Psi_{n\pm})\cdot(\Phi_{1\mp}\wedge\cdots\wedge\Phi_{n\mp}), \tag{2.19}$$

which is a consequence of the following simple algebraic lemma.

LEMMA 2.1. *For all n-tuple of vectors $(W_1,\dots,W_n)$ in $\mathbb{C}^{2n}$ and for all n-tuple of co-vectors $(Z_1,\dots,Z_n)$, we have*

$$\det(Z_i\cdot W_j) = (Z_1\wedge\cdots\wedge Z_n)\cdot(W_1\wedge\cdots\wedge W_n). \tag{2.20}$$

Proof. The proof is immediate since both quantities are $2n$-linear and coincide on the basis of $(\mathbb{C}^{2n*})^n\times(\mathbb{C}^{2n})^n$. ∎

Actually a particular form of (2.19) has already been used by Alexander et al. in [2]. In their simpler forms of (2.17) and (2.18), $D_\pm$ only involve a $n\times n$ determinant instead of a $2n\times 2n$ determinant for the original or dual one. This may simplify the computations, as we shall see below. Moreover, they are likely to generalize to an infinite-dimensional setting. Assume, for example, that the dynamical system (2.3) is infinite-dimensional but that the invariant subspaces $U_\pm(\lambda)$ are finite-dimensional. Then we may define D_+ as in (2.17). This has been used to study the stability of semidiscrete shock profiles (see [3]). Returning to the finite-dimensional framework, the functions $D_\pm$ are closely related to the standard Evans function. More precisely, we have the following result, which generalizes [20, Proposition 1.15].

PROPOSITION 2.2. *There exist analytic functions, $\beta^\pm$, which do not vanish in $\Omega\cup\mathcal{N}$, such that*

$$D(\lambda) = \beta^+(\lambda)D_+(\lambda) = \beta^-(\lambda)D_-(\lambda). \tag{2.21}$$

Proof. We give it only for D_+ since the treatment of D_- is similar. We recall that the standard Evans function is given by

$$D(\lambda) = \det(\Phi_{1-}(0;\lambda), \dots, \Phi_{n-}(0;\lambda), \Phi_{1+}(0;\lambda), \dots, \Phi_{n+}(0;\lambda)). \tag{2.22}$$

There is an analytic choice of solutions $\Phi_{+1}, \dots, \Phi_{+n}$, extending $\{\Phi_{1+}, \dots, \Phi_{n+}\}$, to a basis of solutions to (2.3). Let us define

$$\beta^+(\lambda) := \det(\Phi_{+1}(0;\lambda), \dots, \Phi_{+n}(0;\lambda), \Phi_{1+}(0;\lambda), \dots, \Phi_{n+}(0;\lambda)). \tag{2.23}$$

By construction β^+ does not vanish. There is also a choice of solutions to (2.10), $\Psi_{+1}, \dots, \Psi_{+n}$, such that

$$(\Psi_{1+}, \dots, \Psi_{n+}, \Psi_{+1}, \dots, \Psi_{+n})$$

is a dual basis of

$$(\Phi_{+1}, \dots, \Phi_{+n}, \Phi_{1+}, \dots, \Phi_{n+}),$$

that is to say,

$$\begin{aligned} \Psi_{i+} \cdot \Phi_{j+} = \Psi_{+i} \cdot \Phi_{+j} = 0, \\ \Psi_{+i} \cdot \Phi_{j+} = \Psi_{i+} \cdot \Phi_{+j} = \delta_{ij}. \end{aligned} \tag{2.24}$$

Then the result follows from the definitions (2.17), (2.22), (2.23) and these matrix identities:

$$\begin{pmatrix} \Psi_{1+} \\ \vdots \\ \Psi_{n+} \\ \Psi_{+1} \\ \vdots \\ \Psi_{+n} \end{pmatrix} \begin{pmatrix} \Phi_{+1}, \dots, \Phi_{+n}, \Phi_{1+}, \dots, \Phi_{n+} \end{pmatrix} = I_{2n},$$

which is equivalent to (2.24), and

$$\begin{pmatrix} \Psi_{1+} \\ \vdots \\ \Psi_{n+} \\ \Psi_{+1} \\ \vdots \\ \Psi_{+n} \end{pmatrix} \begin{pmatrix} \Phi_{1-}, \dots, \Phi_{n-}, \Phi_{1+}, \dots, \Phi_{n+} \end{pmatrix} = \begin{pmatrix} \Psi_{i+} \cdot \Phi_{j-} & 0 \\ * & I_n \end{pmatrix}.$$

In particular, we note that the order of vanishing of D and $D_\pm$ at an eigenvalue λ is the same. ∎

3. The framework of viscous shock waves. We now consider a system of viscous conservation laws

$$u_t + f(u)_x = (B(u)u_x)_x, \tag{3.1}$$

and a traveling wave $u(x,t) = U(x-ct)$, tending to asymptotic values $u_\pm$ at $\pm\infty$. Assuming that $B(v)$ is invertible for every $v \in \mathbb{R}^n$, we obtain the ODE satisfied by U,

$$U' = G(U) := B(U)^{-1}(f(U) - f(u_-) - c(U-u_-)), \tag{3.2}$$

as well as the consistency property (Rankine–Hugoniot condition)

$$f(u_+) - f(u_-) = c(u_+ - u_-).$$

Performing a change of variable $(x,t) \mapsto (x-ct,t)$, one may view U as a steady solution of the equation

$$u_t + (f(u) - cu)_y = (B(u)u_y)_y.$$

To study its linear stability, we introduce the linearized equation

$$w_t = Lw := (B(U)w_y + (dB(U)w)U' - (df(U) - cI_n)w)_y. \tag{3.3}$$

The operator L is of the form (2.6) with

$$A(x)\cdot w := (df(U(x)) - cI_n)\cdot w - (dB(U(x))\cdot w)\cdot U'(x), \tag{3.4}$$

and with the abuse of notation

$$B(x) := B(U(x)). \tag{3.5}$$

In particular, A and B have limits at $\pm\infty$, which we shall denote by

$$A_\pm := df(u_\pm) - cI, \quad B_\pm := B(u_\pm).$$

Assuming that the states $u_\pm$ are hyperbolic rest points of G, that is, $dG(u_\pm) = B(u_\pm)^{-1}A_\pm$ do not have pure imaginary eigenvalues, we know that U' converges exponentially fast to zero. Therefore, the linear operator has the same structure as in section 2. We now make standard hypotheses which allow us to apply the gap lemma (see [12]).

(H1) The matrices $A_\pm$ have distinct, nonzero real eigenvalues;
(H2) the eigenvalues of $B(v)$ are of (strictly) positive real part;
(H3) the matrices $B(u_\pm)^{-1}A_\pm$ do not have purely imaginary eigenvalues;
(H4) there exists $\varepsilon > 0$, such that

$$\text{for all } \xi \in \mathbb{R}, \quad \operatorname{Re}\sigma(i\xi A_\pm - \xi^2 B_\pm) \le -\varepsilon\xi^2.$$

Hypothesis (H1) means the strict hyperbolicity of the reduced system

$$u_t + f(u)_x = 0 \tag{3.6}$$

and that the shock wave is noncharacteristic.

The assumption (H4) is known as the Majda–Pego stability condition [19]. It implies that, for $\operatorname{Re}\lambda > 0$, the roots of

$$P^\pm(X;\lambda) := \det(X^2 B_\pm + XA_\pm - \lambda I_n)$$

have a nonzero real part, which is equivalent to the fact that the matrices $\mathbb{A}_\pm(\lambda)$ as defined in (2.9) have no purely imaginary eigenvalues.

Let us point out that (H2)–(H4) are fulfilled whenever

(H5) in the neighborhood of the constant states $u_\pm$, the system of conservation laws (3.1) is consistent with an energy inequality of the form

$$E(u)_t + F(u, u_x)_x + \omega \|u_x\|^2 \leq 0,$$

where $D^2E(u_\pm) > 0_n$ and $\omega > 0$.

Actually, (H5) also implies the hyperbolicity of the reduced system (3.6), though not the strict hyperbolicity.

Remark 2. Most "physical" systems satisfy an energy inequality as stated in (H5) [18, 21].

DEFINITION 3.1. *The viscous shock U is said to be* linearly (spectrally) stable *if* $\sigma(L)$ *lies entirely in the left half-space* $\operatorname{Re}\lambda \leq 0$.

From the above assumptions, this is certainly true regarding the essential part of the spectrum. Therefore, it remains to analyze the possible eigenvalues of positive real part. This is the role of the Evans functions.

We point out that spectral stability is clearly necessary for bounded linearized stability. In fact, the results of [24] give a sufficient condition for linearized (and nonlinear) asymptotic orbital stability in terms of the Evans function, as well, namely, that the only zero on $\{\operatorname{Re}\lambda \geq 0\}$ should lie at $\lambda = 0$, with multiplicity agreeing with the dimension of the manifold of possible viscous profiles; for a complete discussion, see [24].

Regarding $\lambda = 0$, let us note that, by derivation of the profile equation (3.2), one has

$$LU' = 0. \tag{3.7}$$

Since $U' \in L^2$, this means that $\lambda = 0$ is a genuine eigenvalue, and hence

$$D(0) = 0.$$

It has been proved by Zumbrun and Howard [24] that $D_*(0)$ vanishes, too, but we shall give an alternate proof of this fact in section 4.

The strategy of Gardner and Zumbrun in [12] was the following. Pointing out that $D(\lambda)$ is real on $(0, +\infty)$, one may derive an *instability condition* by writing that the sign of D changes between 0^+ and $+\infty$. The sign of D for large real λ comes straightforwardly and involves only the elliptic nature of L, because the first order terms are negligible for high-frequency oscillations; however, it depends on the choice of bases at $\pm\infty$ in the definition of $D(\cdot)$ given in section 2.1. The important contribution of [12], besides the gap lemma, consists in showing that (i) $D'(0)$ may be computed in terms of the hyperbolic structure (see also [25] for a multi-d version of this fact), again up to normalization by choice of bases at $x \to \pm\infty$, and (ii) the choice of bases at $x \to \pm\infty$ made for $\lambda = 0$ can be tracked to $\lambda = +\infty$ to yield a definite sign for $D(\cdot)$; the latter had previously only been carried out in the trivial case of scalar diffusion. Therefore, as soon as $D'(0)$ and $D(\lambda)$ for $\lambda \gg 1$ are of opposite sign, the intermediate value theorem ensures the existence of a positive real eigenvalue for L, which in turn means that U is unstable. We shall show in sections 4 and 5 that a similar approach can be used concerning the alternate Evans functions defined in section 2. A notable limitation in [12] was that (ii) was carried out only for $n = 2$, though the procedure used was conjectured to remain valid in the general case; the resolution of this issue was emphasized as a key open problem in the theory. In section 7, we verify the conjecture of [12], thus completing the theory for $n > 2$.

Let us fix some notations. The eigenvalues of $A_\pm$ are denoted by $a_j^\pm$ $(1 \le j \le n)$. These are distinct nonzero real numbers, that we order by

$$a_1^\pm < \cdots < a_n^\pm.$$

Let us remark that the existence of the viscous shock U implies that $a_1^+ < 0 < a_n^-$. We shall denote by $(r_1^\pm, \dots, r_n^\pm)$ a corresponding basis of right eigenvectors and by $(\ell_1^+, \dots, \ell_n^+)$ a basis of left eigenvectors. We shall feel free to choose their orientations, in particular, according to the profile. By assumption (H3), the eigenvalues $\gamma_j^\pm$ $(1 \le j \le n$, counting with multiplicities) of $B_\pm^{-1}A_\pm$ have a nonzero real part. Furthermore, a continuation argument shows the following (see [12, Lemma 3.8]).

LEMMA 3.2. *Assuming* (H1) *through* (H4)*, the unstable/stable manifolds of* $A_\pm$ *and* $B_\pm^{-1}A_\pm$ *have equal dimensions.*

Therefore, we may assume that $\operatorname{Re}\gamma_j^\pm$ has the same sign as $a_j^\pm$.

Remark 3. Lemma 3.2 does not require the strict hyperbolicity of $A_\pm$. It also holds with (H1) replaced by

(H1′) the matrices $A_\pm$ are invertible and diagonalizable on $\mathbb{R}$.

This will be used in section 7.

4. Viscous shock waves via the dual Evans function. The main purpose of this section is to present a computation of $D_*'(0)$, which differs significantly from the one of $D'(0)$. In addition, we perform the analysis of D_* near $+\infty$ by means of a homotopy argument, which differs from the rescaling method used in [12]. As expected, we shall find the same instability criterion as in [12], since they both tell that the number of positive real eigenvalues of L (respectively, L^*) are odd, and we already know that these numbers are equal to each other.

Let us recall that, when $\operatorname{Re}\lambda > 0$, the assumption (H4) ensures that the roots of

$$P^\pm(X;\lambda) = \det(X^2B_\pm + XA_\pm - \lambda I_n)$$

have a nonzero real part. By a continuity argument, we find that n of them have a positive real part, while the n others have a negative real part. We select those of $P^+(\cdot;\lambda)$ which have a negative real part, and call them $\mu_j^+(\lambda)$ $(j = 1, \dots, n)$; we make the opposite choice for P^-. In short,

$$\det\left((\mu_j^\pm)^2B_\pm + \mu_j^\pm A_\pm - \lambda I_n\right) = 0, \quad \pm\operatorname{Re}\mu_j^\pm < 0.$$

We shall denote by $y_{j\pm}$ a corresponding left "eigenvector" such that

$$y_{j\pm}(\lambda)\left((\mu_j^\pm)^2B_\pm + \mu_j^\pm A_\pm - \lambda I_n\right) = 0.$$

By standard matrix theory, these vectors can be chosen analytically in regions where the $\mu_j^\pm$ do not "cross" each other.

Let us first concentrate on the long wave analysis, that is, the limit $\lambda \to 0^+$. We first describe the behavior of the $\mu_j^\pm$. Clearly, the limits of all the roots of $P^\pm$ are on one hand the eigenvalues of $-A_\pm B_\pm^{-1}$, which provide n roots, counting with multiplicities, and on the other hand $\mu = 0$, with multiplicity n. Actually, the small roots are equivalent to $\lambda/a_j^\pm$. We conclude that

$$\mu_j^+(\lambda) \sim \begin{cases} -\gamma_j^+ & \text{if } a_j^+ > 0, \\ \lambda/a_j^+ & \text{if } a_j^+ < 0. \end{cases}$$

Similarly,

$$\mu_j^-(\lambda) \sim \begin{cases} \lambda/a_j^- & \text{if } a_j^- > 0, \\ -\gamma_j^- & \text{if } a_j^- < 0. \end{cases}$$

As for the operator L, we choose bases of decaying solutions of the ODE $L^*\chi = \lambda\chi$, as $\operatorname{Re}\lambda > 0$. These bases are chosen holomorphically. They consist of n functions denoted by $\{\chi_{1\pm}(x;\lambda), \dots, \chi_{n\pm}(x;\lambda)\}$. These bases are analytically extended to a neighborhood of the origin, but then $\chi_{j\pm}(\lambda)$ need not decay as $x \to \pm\infty$. However, they satisfy

$$\chi_{j\pm}(x;\lambda) \underset{x \to \pm\infty}{\sim} e^{\mu_j^\pm(\lambda)x}\, y_{j\pm}(\lambda), \tag{4.1}$$

when λ is close enough to 0 (see [12, Proposition 3.2]). More precisely, the behavior of the basis functions at $\lambda = 0$ can be summarized as follows.

- When $\pm a_j^\pm > 0$, then
$$\mu_j^\pm(0) = -\gamma_j^\pm$$
and we still have $\pm\operatorname{Re}\mu_j^\pm < 0$. Thus
$$\chi_{j\pm}(\pm\infty;0) = 0,$$
with exponential decay towards the direction
$$y_{j\pm}(0) = k_j^\pm,$$
where $k_j^\pm$ is a left eigenvector of $A_\pm B_\pm^{-1}$ associated to $\gamma_j^\pm$, provided that this matrix is diagonalizable.
- When $\pm a_j^\pm < 0$, then
$$\mu_j^\pm(0) = 0$$
and by a standard bifurcation argument we find that
$$y_{j\pm}(0) = \ell_j^\pm,$$
where $\ell_j^\pm$ is a left eigenvector of $A_\pm$ associated to $a_j^\pm$. Actually, $\chi_{j\pm}$ is then *constant*:
$$\chi_{j\pm}(\cdot;0) \equiv \ell_j^\pm.$$
In particular, (4.1) trivially holds.

We shall also be concerned with the λ-derivatives

$$\theta_{j\pm}(x) := \frac{\partial\chi_{j\pm}}{\partial\lambda}(x;0).$$

These functions satisfy a nonhomogeneous equation $L^*\theta = \chi$. Moreover, when $\pm a_j^\pm > 0$ we should have

$$\theta_{j\pm}(\pm\infty;0) = 0.$$

The dual Evans function is defined by (2.8).

In the remainder of this section, we shall restrict our discussion to 2×2 systems. We investigate the three possible cases: a Lax shock, an under-compressive shock, and an over-compressive shock. In the latter, we have $D'(0) = D(0) = 0$, so that the sign of $D(0^+)$ is determined through $D''(0)$, and similarly for D_*. These cases are characterized according to the signs of the $a_j^\pm$:

- a 1-Lax shock:

$$a_1^+ < 0 < a_1^-, a_2^+, a_2^-;$$

- a 2-Lax shock:

$$a_1^+, a_1^-, a_2^+ < 0 < a_2^-;$$

- an under-compressive shock:

$$a_1^+, a_1^- < 0 < a_2^+, a_2^-;$$

- an over-compressive shock:

$$a_1^+, a_2^+ < 0 < a_1^-, a_2^-.$$

4.1. The Lax shock case. By symmetry we need only to investigate a 2-Lax shock. We have $a_1^\pm, a_2^+ < 0 < a_2^-$ and similarly for the $\gamma_j^\pm$. It follows that, for $\lambda = 0$, we have

$$\chi_{1-}(-\infty) = 0, \quad \chi_{2-} \equiv \ell_2^-, \quad \chi_{1+} \equiv \ell_1^+, \quad \chi_{2+} \equiv \ell_2^+. \tag{4.2}$$

We easily conclude that

$$D_* = \left| \begin{array}{cc} \chi_{1-} & \chi'_{1-} \\ \chi_{2-} & \chi'_{2-} \\ \chi_{1+} & \chi'_{1+} \\ \chi_{2+} & \chi'_{2+} \end{array} \right|_{x=0}$$

vanishes at $\lambda = 0$, since the two last columns of the determinant are linearly dependent, each being of the form $(\cdot, 0, 0, 0)^T$.

We now compute[2] $D'_*(0)$. It is the sum of four determinants, each one obtained from the former by replacing a row (χ, χ') by the corresponding co-vector (θ, θ'). The first one vanishes for the same reason as before, while the three others are block-triangular. For instance,

$$\left| \begin{array}{cc} \chi_{1-} & \chi'_{1-} \\ \theta_{2-} & \theta'_{2-} \\ \chi_{1+} & \chi'_{1+} \\ \chi_{2+} & \chi'_{2+} \end{array} \right| = \left| \begin{array}{cc} \chi_{1-} & \chi'_{1-} \\ \theta_{2-} & \theta'_{2-} \\ \ell_1^+ & 0 \\ \ell_2^+ & 0 \end{array} \right| = (\chi'_{1-} \wedge \theta'_{2-})(\ell_1^+ \wedge \ell_2^+), \tag{4.3}$$

where the wedge products are identified to their component on the canonical 2-form, that is,

$$\chi \wedge \kappa = \left| \begin{array}{c} \chi \\ \kappa \end{array} \right|.$$

Thus we obtain the formula

$$D'_*(0) = (\chi'_{1-} \wedge \theta'_{2-})(\ell_1^+ \wedge \ell_2^+) + (\chi'_{1-} \wedge \theta'_{1+})(\ell_2^+ \wedge \ell_2^-) + (\chi'_{1-} \wedge \theta'_{2+})(\ell_2^- \wedge \ell_1^+), \tag{4.4}$$

[2]We hope that the prime, which sometimes denotes a λ-derivative and elsewhere denotes an x-derivative, will not give rise to confusion.

all terms being taken at $x = 0$.

Now let us give another form of the wedge products $(\chi'_{1-} \wedge X)$. We have the following result.

LEMMA 4.1. *There exists a smooth scalar function c on $\mathbb{R}$ such that, for all co-vector X, for all $x \in \mathbb{R}$*

$$\chi'_{1-}(x) \wedge X = c(x) X B(x) U'(x). \tag{4.5}$$

Moreover, c satisfies the differential equation

$$c' + \operatorname{tr}\left((B' + A)B^{-1}\right)c = 0. \tag{4.6}$$

Proof. Since by (3.7) and decay assumptions the profile U is such that

$$BU'' - A \cdot U' \equiv 0, \tag{4.7}$$

we have, for any solution z to $L^* z = (z'B)' + z'A = 0$, that

$$(z'BU')' \equiv 0.$$

Hence, in particular, $\chi'_{1-} BU'$ is a constant. This constant is zero because χ'_{1-} (as well as U') vanishes at $-\infty$. Since U' does not vanish in $\mathbb{R}$, there must exist a function c such that $\chi'_{1-} \wedge X = cXBU'$. We now differentiate this identity

$$c'XBU' + cX(BU')' = \chi''_{1-} \wedge X.$$

On one hand, we have from (4.7) that

$$(BU')' = (B' + A)U',$$

and on the other hand, we have

$$\begin{aligned} \chi''_{1-} \wedge X &= -\chi'_{1-}(B' + A)B^{-1} \wedge X \\ &= -\operatorname{tr}((B' + A)B^{-1})(\chi'_{1-} \wedge X) + \chi'_{1-} \wedge X(B' + A)B^{-1} \\ &= -\operatorname{tr}((B' + A)B^{-1})(\chi'_{1-} \wedge X) + cX(B' + A)U', \end{aligned}$$

by the definition (4.5). Thus we find that for all X

$$\left(c' + \operatorname{tr}((B' + A)B^{-1})c\right) XBU' = 0.$$

This proves the formula, because U' does not vanish. ∎

Let us put the identity (4.5) into (4.4):

$$\begin{aligned} c(0)^{-1} D'_*(0) = {} & (\theta'_{2-} BU')(\ell_1^+ \wedge \ell_2^+) + (\theta'_{1+} BU')(\ell_2^+ \wedge \ell_2^-) \\ & + (\theta'_{2+} BU')(\ell_2^- \wedge \ell_1^+). \end{aligned} \tag{4.8}$$

Now we compute the terms of the form $\theta' BU'$. Since in all three cases we have $L^* \theta_{j\pm} = \chi_{j\pm} \equiv \ell_j^\pm$, we find using (4.7) that

$$(\theta'_{j\pm} BU')' = \ell_j^\pm U'.$$

This yields $\theta'_{j\pm} BU' = \ell_j^\pm (U - \text{constant})$. The constant is computed by letting $x \to \pm\infty$, since $\theta_{j\pm}$ is at most algebraically growing near $\pm\infty$, whereas U' decays

exponentially to zero on both sides. Therefore, the right-hand side must vanish at $\pm\infty$, which provides

$$\theta'_{j\pm} B U' = \ell_j^{\pm}(U - u_{\pm}).$$

This leads to

$$\begin{aligned} c(0)^{-1} D'_*(0) = {} & \ell_2^-(U-u_-)(\ell_1^+ \wedge \ell_2^+) + \ell_1^+(U-u_+)(\ell_2^+ \wedge \ell_2^-) \\ & + \ell_2^+(U-u_+)(\ell_2^- \wedge \ell_1^+). \end{aligned} \tag{4.9}$$

However, we have for all co-vectors X, Y, Z the identity

$$(X \wedge Y)\, Z + (Z \wedge X)\, Y + (Y \wedge Z)\, X = 0.$$

Therefore, (4.9) simplifies into

$$\begin{aligned} c(0)^{-1} D'_*(0) &= -\ell_2^- u_-(\ell_1^+ \wedge \ell_2^+) - \ell_1^+ u_+(\ell_2^+ \wedge \ell_2^-) - \ell_2^+ u_+(\ell_2^- \wedge \ell_1^+) \\ &= \ell_2^-[u](\ell_1^+ \wedge \ell_2^+). \end{aligned}$$

At this stage, it remains to compute the sign of $c(0)$. Actually, c has a constant sign, so that it is sufficient to evaluate it as $x \to -\infty$. We recall that

$$U'(x) \underset{x \to -\infty}{\sim} e^{\gamma_2^- x}\, s_2^-, \tag{4.10}$$

where $s_j^{\pm}$ is a right eigenvector of $B_{\pm}^{-1} A_{\pm}$ corresponding to the eigenvalue $\gamma_j^{\pm}$. (Let us point out that γ_1^- and γ_2^- cannot be conjugate to each other and thus must be real; they are simple eigenvalues of $A_- B_-^{-1}$ as well as of $B_-^{-1} A_-$.) Similarly,

$$\chi'_{1-}(x) \underset{x \to -\infty}{\sim} e^{-\gamma_1^- x}\, k_1^-, \tag{4.11}$$

where k_j^- is a left eigenvector of $A_- B_-^{-1}$ associated to γ_j^-. We point out that

$$k_j^- A_- s_k^- = k_j^- B_- s_k^- = 0$$

when $j \neq k$, whereas

$$k_j^- B_- s_j^- \neq 0.$$

We obtain from the definition (4.5) and the asymptotic behaviors (4.10) and (4.11) that

$$c(x) \underset{x \to -\infty}{\sim} \frac{(k_1^- \wedge X)}{X B_- s_2^-}\, e^{-\mathrm{tr}(A_- B_-^{-1})x}$$

for every co-vector X not parallel to k_1^-. For instance,

$$c(x) \underset{x \to -\infty}{\sim} \frac{(k_1^- \wedge k_2^-)}{k_2^- B_- s_2^-}\, e^{-\mathrm{tr}(A_- B_-^{-1})x}.$$

Finally,

$$\mathrm{sgn}\ D'_*(0) = \mathrm{sgn}\ (\ell_1^+ \wedge \ell_2^+)(k_1^- \wedge k_2^-)(k_2^- B_- s_2^-)\ \ell_2^-[u]. \tag{4.12}$$

In the above formula, several vectors are known up to a real factor and might be reversed by a new choice of the basis functions $k_1^-, \ell_2^-, \ell_j^+$. We do not worry with the orientation of k_2^-, which occurs twice. Besides, s_2^- may not be reversed, because it is entirely determined by the profile itself. Changes in the four vectors induce a possible change of the sign of the Evans function, which reflects in a change of the signs of $D_*'(0)$ and of $D_*(+\infty)$ simultaneously. Therefore, we have to express the sign of $D_*(+\infty)$ in terms of these vectors in order to know modulo 2 the number of positive real eigenvalues of L^*.

4.2. The sign of $D_*(\lambda)$ for large real λ. We present here a method that differs from the one employed in [12] and which may be simpler. It is versatile and can be used even in the evaluation of the direct Evans function. We first build a *homotopy*, for $\lambda > \Lambda$, between $\lambda \mapsto L^* - \lambda$ and $d^2/dx^2 - \lambda$:

$$L_\theta^* := \theta L^* + (1-\theta)\frac{d^2}{dx^2}, \quad \theta \in [0,1].$$

Here Λ is a real number, large enough to ensure that $L_\theta^* - \lambda$ is invertible for $\lambda > \Lambda$ and $\theta \in [0,1]$. Such a number exists because of the Gårding inequality: there exist two numbers $\omega > 0$ and Λ such that

$$(L^* z | z) + \omega \|z_x\|_{L^2} \le \Lambda \|z\|_{L^2}.$$

As in [12], but adding θ to the parameter λ, we may choose the basis functions $\chi_j^\pm$ depending analytically both on λ and θ. Thus we define a two-parameters Evans function $D_*(\lambda,\theta)$, which is real-valued if λ is real. When $\lambda > \Lambda$ and $\theta \in [0,1]$, D_* does not vanish. Therefore, the sign of $D_*(\lambda) = D_*(\lambda,1)$, for large λ, is the same as the sign of $D_*(\lambda,0)$. The latter may be computed easily. When $\theta = 0$, the eigenfunctions have the form

$$\chi_j^\pm(x;\lambda,0) = e^{\mp x\sqrt{\lambda}}\, V_j^\pm,$$

where $V_j^\pm$ are some real co-vectors. Thus

$$D_*(\lambda,0) = \begin{vmatrix} V_1^- & \sqrt{\lambda}V_1^- \\ V_2^- & \sqrt{\lambda}V_2^- \\ V_1^+ & -\sqrt{\lambda}V_1^+ \\ V_2^+ & -\sqrt{\lambda}V_2^+ \end{vmatrix} = 4\lambda (V_1^- \wedge V_2^-)(V_1^+ \wedge V_2^+).$$

Now it follows from a generalized version of Lemma 3.5 in [12], involving the additional parameter θ, that by continuity

$$\begin{array}{rcl} \operatorname{sgn}\,(V_1^- \wedge V_2^-) & = & \operatorname{sgn}\,(y_1^-(0) \wedge y_2^-(0)), \\ \operatorname{sgn}\,(V_1^+ \wedge V_2^+) & = & \operatorname{sgn}\,(y_1^+(0) \wedge y_2^+(0)). \end{array} \tag{4.13}$$

Since we have

$$y_j^+(0) = \ell_j^+,\ y_2^-(0) = \ell_2^-,$$

and, in view of (4.1) and (4.11),

$$y_1^-(0) = -k_1^-/\gamma_1^-,$$

with $-\gamma_1^- > 0$, (4.13) yield

$$\operatorname{sgn} D_*(\lambda) = \operatorname{sgn} (k_1^- \wedge \ell_2^-)(\ell_1^+ \wedge \ell_2^+) \tag{4.14}$$

for large real λ. Finally, we find that

$$\operatorname{sgn} D_*'(0)\, D_*(+\infty) = \operatorname{sgn} (k_1^- \wedge \ell_2^-)(k_1^- \wedge k_2^-)(k_2^- B_- s_2^-)\, \ell_2^-[u]. \tag{4.15}$$

Let us recall from [12] that the result for the standard Evans function is

$$\operatorname{sgn} D'(0)\, D(+\infty) = \operatorname{sgn} (r_1^- \wedge s_2^-)(r_1^- \wedge [u]). \tag{4.16}$$

It is not difficult to check from (4.15) and (4.16) that

$$\operatorname{sgn} D_*'(0)\, D_*(+\infty) = \operatorname{sgn} D'(0)\, D(+\infty).$$

As a matter of fact, let us assume for simplicity that the ℓ_i and k_i are normalized by

$$\ell_i r_j = \delta_{ij}$$

and

$$k_i B s_j = \delta_{ij}.$$

Then we have

$$\begin{aligned} r_1^- \wedge [u] &= \ell_2^-[u](r_1^- \wedge r_2^-), \\ r_1^- \wedge s_2^- &= \ell_2^- s_2^-(r_1^- \wedge r_2^-). \end{aligned}$$

Moreover, we have

$$\begin{pmatrix} k_1^- \\ k_2^- \end{pmatrix} B_- \begin{pmatrix} s_1^- & s_2^- \end{pmatrix} = I_2,$$

and thus

$$\operatorname{sgn} (k_1^- \wedge k_2^-) = \operatorname{sgn} (s_1^- \wedge s_2^-),$$

since $\det B_- > 0$. Similarly, the identity

$$\begin{pmatrix} k_1^- \\ \ell_2^- \end{pmatrix} A_- \begin{pmatrix} s_1^- & s_2^- \end{pmatrix} = \begin{pmatrix} \gamma_1^- & 0 \\ * & a_2^- \ell_2^- s_2^- \end{pmatrix}$$

implies

$$\operatorname{sgn} (k_1^- \wedge \ell_2^-) = \operatorname{sgn} (s_1^- \wedge s_2^-)\, \ell_2^- s_2^-.$$

4.3. The case of an under-compressive shock. In this paragraph, as well as in the next one, we focus only on the computation of $D_*'(0)$ (or $D_*''(0)$ if needed) once arbitrary choices of the eigenforms $\ell_j^\pm$ are made. The computation of the sign of D_* near infinity follows the same lines as in the Lax shock case. It splits into terms associated to the unstable/stable manifold only. Each term can be followed from large real λ to $\lambda = 0$ in the same way as in [12] (if $n = 2$) or in section 7 below (for general n).

We have here $a_1^\pm < 0 < a_2^\pm$ and similarly for the $\gamma_j^\pm$. It follows that, for $\lambda = 0$, we have

$$\chi_{1-}(-\infty) = 0, \quad \chi_{2-} \equiv \ell_2^-, \quad \chi_{1+} \equiv \ell_1^+, \quad \chi_{2+}(+\infty) = 0. \tag{4.17}$$

From this, we see that $D_*(\lambda = 0)$ is block-triangular, so that

$$D_*(0) = -(\ell_1^+ \wedge \ell_2^-)(\chi_{1-}' \wedge \chi_{2+}').$$

Using Lemma 4.1, which is still valid with our assumptions, we obtain

$$\chi_{1-}' \wedge \chi_{2+}' = c\chi_{2+}' BU'.$$

However, the same argument as the one in the proof of the lemma gives $\chi_{2+}' BU' \equiv 0$. Therefore, D_* vanishes at the origin, as expected, and we need to calculate $D_*'(0)$. Before doing so, we remark that we proved $\chi_{1-}' \wedge \chi_{2+}' = 0$. Since they both solve the same first order ODE $(\xi B)' + \xi A = 0$, this means (by uniqueness for the Cauchy problem) that $\chi_{2+}' = C\chi_{1-}'$, where C is a constant. There is no loss of generality to fix $C = 1$. We assume from now on that $\ell_1^+ \wedge \ell_2^- \neq 0$, so that $\chi_{2+} - \chi_{1-}$, being constant, is a linear combination of χ_{1+} and χ_{2-}.

Again $D_*'(0)$ is the sum of four determinants, each one obtained from $D_*(0)$ by replacing a row (χ, χ') by the corresponding (θ, θ'). Two of them vanish because the last row is a combination of the first three. The two others are block-triangular. Therefore, denoting χ_{1-} by χ, which equals χ_{2+} modulo constants,

$$D_*'(0) = -(\ell_1^+ \wedge \ell_2^-)\left(\chi' \wedge (\theta_{2+}' - \theta_{1-}')\right).$$

Using Lemma 4.1, this reduces to

$$D_*'(0) = -c(\ell_1^+ \wedge \ell_2^-)(\theta_{2+}' - \theta_{1-}')BU',$$

with everything being computed at $x = 0$.

However, $(\theta_{1-}' BU')' = \chi \cdot U'$ and $\theta_{1-}' BU'(-\infty) = 0$ give

$$\theta_{1-}' BU'(0) = \int_{-\infty}^0 \chi \cdot U' dx,$$

and similarly

$$\theta_{2+}' BU'(0) = -\int_0^{+\infty} \chi \cdot U' dx.$$

Finally,

$$D_*'(0) = c(0)(\ell_1^+ \wedge \ell_2^-) \int_{-\infty}^{+\infty} \chi \cdot U' dx. \tag{4.18}$$

Remarks.

- We recognize the Melnikov integral (see (3.18) in [12]) of Gardner and Zumbrun in the second factor of the right-hand side.
- The formula (4.18) shows that the assumption $\ell_1^+ \wedge \ell_2^- \neq 0$ is essential in the analysis, equivalent to assumption $r_1^- \wedge r_2^+ \neq 0$ in [12]. When this condition fails, a different calculation (in similar spirit) must be carried out; see [12].

- The last factor in the right-hand side of (4.18) does not depend on the choice of χ, modulo constants. In other words, one might have chosen $\chi = \chi_{2+}$.
- For larger systems ($n \geq 2$), with assumption $a_p(u_\pm) < 0 < a_{p+1}(u_\pm)$, we obtain the similar formula

$$D'_*(0) = c(\ell_1^+ \wedge \cdots \wedge \ell_p^+ \wedge \ell_{p+1}^- \wedge \cdots \wedge \ell_n^-) \int_{-\infty}^{+\infty} \chi \cdot U' dx,$$

where χ is a nonconstant bounded solution of $L^*\chi = 0$. The vector space of bounded solutions is generically of dimension $n+1$ and contains the constants. Besides, c is a nonvanishing function such that

$$\chi' \wedge \chi'_{2-} \wedge \cdots \wedge \chi'_{n-1,+} \wedge X = cXBU'.$$

As usual, c solves the ODE (4.6). Here, we have assumed, similarly as in the case $n = 2$, that

$$\ell_1^+ \wedge \cdots \wedge \ell_p^+ \wedge \ell_{p+1}^- \wedge \cdots \wedge \ell_n^- \neq 0.$$

4.4. The case of an over-compressive shock. Here $a_{1,2}^+ < 0 < a_{1,2}^-$. It follows that, for $\lambda = 0$, we have $\gamma_j^\pm(0) = 0$, so that $\chi_{j\pm} \equiv \ell_j^\pm$. Therefore, $D_*(0) = D'_*(0) = 0$ come trivially. Now, $\frac{1}{2}D''_*(0)$ is the sum of six determinants, each one obtained from D_* by replacing two rows (χ, χ') by the corresponding (θ, θ'). They all are block-triangular. Therefore, we obtain

$$\begin{aligned}\frac{1}{2}D''_*(0) = {} & (\theta'_{1-} \wedge \theta'_{2-})(\ell_1^+ \wedge \ell_2^+) - (\theta'_{1-} \wedge \theta'_{1+})(\ell_2^- \wedge \ell_2^+) \\ & + (\theta'_{1-} \wedge \theta'_{2+})(\ell_2^- \wedge \ell_1^+) + (\theta'_{2-} \wedge \theta'_{1+})(\ell_1^- \wedge \ell_2^+) \\ & - (\theta'_{2-} \wedge \theta'_{2+})(\ell_1^- \wedge \ell_1^+) + (\theta'_{1+} \wedge \theta'_{2+})(\ell_1^- \wedge \ell_2^-).\end{aligned}$$

Now let us notice that our viscous shock U is one among a one-parameter family $\delta \mapsto u^\delta$, since u_- is a repeller and u_+ is an attractor for the dynamical system (3.2). The operator L, therefore, has a double eigenvalue $\lambda = 0$, corresponding to the eigenfunctions U' and

$$v := \left.\frac{du^\delta}{d\delta}\right|_{\delta=0}.$$

Let us multiply $D''_*(0)$ by the nonzero determinant $Bv \wedge BU'$, using the formula

$$(\alpha \wedge \beta)(X \wedge Y) = (\alpha \cdot X)(\beta \cdot Y) - (\alpha \cdot Y)(\beta \cdot X).$$

We obtain a formula such as

$$\frac{\det B}{2}(v \wedge U')D''_*(0) = \text{a sum of 12 products.}$$

Each such term is a product of the form $\pm(\ell \wedge \ell)(\theta' BU')(\theta' Bv)$, with various indices. Let us have a look at the coefficient of $\theta'_{1-}Bv$, for instance. It appears as

$$\begin{aligned}& (\ell_1^+ \wedge \ell_2^+)(\theta'_{2-}BU') + (\ell_2^+ \wedge \ell_2^-)(\theta'_{1+}BU') + (\ell_2^- \wedge \ell_1^+)(\theta'_{2+}BU') \\ &= (\ell_1^+ \wedge \ell_2^+)\ell_2^- \cdot (U - u_-) + (\ell_2^+ \wedge \ell_2^-)\ell_1^+ \cdot (U - u_+) + (\ell_2^- \wedge \ell_1^+)\ell_2^+ \cdot (U - u_+) \\ &= (\ell_1^+ \wedge \ell_2^+)\ell_2^- \cdot [u].\end{aligned}$$

Finally,

$$\frac{\det B}{2}(v\wedge U')D_*''(0) = \theta_{1-}'Bv(\ell_1^+\wedge\ell_2^+)\ell_2^-\cdot[u] + \theta_{2-}'Bv(\ell_2^+\wedge\ell_1^+)\ell_1^-\cdot[u] + \theta_{1+}'Bv(\ell_2^-\wedge\ell_1^-)\ell_2^+\cdot[u] + \theta_{2+}'Bv(\ell_1^-\wedge\ell_2^-)\ell_1^+\cdot[u],$$

where everything is computed at $x=0$. Now we point out that $(\theta_{j\pm}'Bv)' = \ell_j^\pm\cdot v$, and that $\theta_{j\pm}'$ is bounded by a polynomial near $\pm\infty$ while v decays exponentially fast to zero. This ensures

$$\theta_{j\pm}'Bv(\pm\infty) = 0,$$

so that

$$\theta_{j\pm}'Bv(0) = \ell_j^\pm\cdot m_\pm,\quad m_\pm := \int_{\pm\infty}^0 v(x)dx.$$

Therefore,

$$\begin{aligned}\frac{\det B}{2}(v\wedge U')D_*''(0) &= (\ell_1^+\wedge\ell_2^+)(\ell_1^-\cdot m_-\ell_2^-\cdot[u]-\ell_2^-\cdot m_-\ell_1^-\cdot[u])\\ &\quad-(\ell_1^-\wedge\ell_2^-)(\ell_1^+\cdot m_+\ell_2^+\cdot[u]-\ell_2^+\cdot m_+\ell_1^+\cdot[u])\\ &= (\ell_1^-\wedge\ell_2^-)(\ell_1^+\wedge\ell_2^+)(m\wedge[u]),\end{aligned}$$

where

$$m := m_- - m_+ = \int_{-\infty}^{+\infty} v(x)dx.$$

Thus we have recovered the result of [12] in this case, too. Let us remark that the sign of $D_*''(0)$ is completely determined by those of $v\wedge U'$ (constant sign along the real line) and of $m\wedge[u] = \int v\wedge\int U'$. Let us emphasize that these need not be the same. Examples with distinct signs are given in [9].

5. Viscous shock waves via the mixed Evans function. In this section, we show how to derive a stability condition by means of the mixed Evans functions introduced in section 2.3. Our main purpose is to compute the signs of $D_\pm{}'(0)$ and $D_\pm(+\infty)$ in the Lax shock case. We choose to deal with D_+, the treatment of D_- being symmetric.

5.1. The general *p*-shock case. We recall that under the assumption (H1), the stationary discontinuity (u_-,u_+) is a p-shock for some $p\in\{1,\dots,n\}$, if the following inequalities hold:

$$a_1^1 < \cdots < a_{p-1}^- < 0 < a_p^- < \cdots < a_n^-, \tag{5.1}$$

$$a_1^+ < \cdots < a_p^+ < 0 < a_{p+1}^+ < \cdots < a_n^+. \tag{5.2}$$

Assuming the discontinuity is a p-shock, we want to evaluate $D_+{}'(0)$. From the right-hand inequalities in (5.1), it is not difficult to show that for $\lambda=0$ there are still $n-p+1$ independent solutions of (2.3), say, $\Phi_{p-},\dots,\Phi_{n-}$, decaying exponentially at $-\infty$. On the other hand, from the left-hand inequalities in (5.1), we see that $p-1$

decaying solutions for $\operatorname{Re}\lambda > 0$ bifurcate to asymptotically constant solutions for $\lambda = 0$. More precisely, we may choose $\Phi_{1-}, \dots, \Phi_{p-1-}$ such that

$$\Phi_{j-}(x;0) \underset{x \to -\infty}{\longrightarrow} \begin{pmatrix} r_j^- \\ 0 \end{pmatrix}, \quad j = 1, \dots, p-1.$$

In a similar way, we find from the inequalities in (5.2) that there are p independent solutions of (2.3), say, $\Phi_{1+}, \dots, \Phi_{p+}$, decaying exponentially at $+\infty$, and $n-p$ asymptotically constant solutions

$$\Phi_{j+}(x;0) \underset{x \to +\infty}{\longrightarrow} \begin{pmatrix} r_j^+ \\ 0 \end{pmatrix}, \quad j = p+1, \dots, n.$$

In particular, we may choose in view of (3.7)

$$\Phi_{p-} = \Phi_{p+} = \begin{pmatrix} U' \\ U'' \end{pmatrix} \quad \text{at } \lambda = 0.$$

Now, concerning the adjoint equation, the inequalities (5.2) show that there are $n-p$ independent solutions of (2.10), say, $\Psi_{p+1+}, \dots, \Psi_{n+}$, decaying exponentially at $+\infty$, and p asymptotically constant solutions, $\Psi_{1+}, \dots, \Psi_{p+}$. The point is that we can choose the Ψ_{i+} in a very simple way. This is due to the following classical result (see [24, Lemma 4.4], for instance), which relates the adjoint ODE (2.10) to the adjoint operator L^*.

LEMMA 5.1. *Equation* (2.10) *is equivalent to*

$$Z = (z, z')S, \tag{5.3}$$

$$L^* z = \bar{\lambda} z, \tag{5.4}$$

where S is the (variable coefficients) invertible matrix defined by

$$S = \begin{pmatrix} -A & B \\ -B & 0 \end{pmatrix}. \tag{5.5}$$

We recall from (2.15) that we must have

$$\Psi_{i+} \cdot \Phi_{j+} = 0, \quad i, j \in \{1, \dots, n\}. \tag{5.6}$$

This is trivial for the decaying solutions $\Psi_{p+1+}, \dots, \Psi_{n+}$ since the Φ_{j+} are bounded at $+\infty$. But we have to choose $\Psi_{1+}, \dots, \Psi_{p+}$ in order to satisfy (5.6). This is done in the following proposition.

PROPOSITION 5.2. *We assume the transversality condition*

$$\operatorname{Span}\{\ell_1^+, \dots, \ell_p^+, \ell_p^-, \dots, \ell_n^-\} = \mathbb{R}^n. \tag{5.7}$$

We take p independent co-vectors h_i such that

$$h_p \in \operatorname{Span}\{\ell_1^+, \dots, \ell_p^+\} \cap \operatorname{Span}\{\ell_p^-, \dots, \ell_n^-\}, \tag{5.8}$$

$$h_1, \dots, h_{p-1} \in \operatorname{Span}\{\ell_1^+, \dots, \ell_p^+\}. \tag{5.9}$$

Then the bounded functions

$$\Psi_{i+} := (h_i, 0)S = (-h_i A, h_i B), \quad i \in \{1, \dots, p\}, \tag{5.10}$$

are independent solutions of (2.10) *for* $\lambda = 0$ *which satisfy* (5.6).

Proof. The existence of h_p as in (5.8) follows from the condition (5.7) and the independence of $\ell_1^+, \dots, \ell_p^+$, as well as of the independence of $\ell_p^-, \dots, \ell_n^-$. Since the constants belong to the kernel of L^*, the fact that the Ψ_{i+} as defined in (5.10) are solution to (2.10) for $\lambda = 0$ is a straightforward consequence of Lemma 5.1. Furthermore, we have for all $j \in \{1, \dots, p\}$

$$\Psi_{i+} \cdot \Phi_{j+} = 0,$$

since Φ_{j+} vanishes at $+\infty$, and for all $j \in \{p+1, \dots, n\}$

$$\Psi_{i+} \cdot \Phi_{j+} = -h_i A_+ r_j^+ = -a_j^+ h_i r_j^+ = 0,$$

by evaluating the (constant) dot product at $+\infty$ and using the fact that $h_i \in \mathrm{Span}\{\ell_1^+, \dots, \ell_p^+\}$. ∎

The particular choice of Ψ_{p+} is meant to cancel as many terms as possible in $D_+{}'(0)$. As a matter of fact, it implies that

$$\Psi_{p+} \cdot \Phi_{j-} = 0, \quad j \in \{1, \dots, n\}. \tag{5.11}$$

It is straightforward if $j \in \{p, \dots, n\}$ since the Φ_{j-} vanish at $-\infty$, and if $j \in \{1, \dots, p-1\}$, we have

$$\Psi_{p+} \cdot \Phi_{j-} = -h_p A_- r_j^- = -a_j^- h_p r_j^- = 0,$$

since $h_p \in \mathrm{Span}\{\ell_p^-, \dots, \ell_n^-\}$. Moreover, we recall from the choice of Φ_{p-} that

$$\Psi_{i+} \cdot \Phi_{p-} = 0, \quad i \in \{1, \dots, n\}. \tag{5.12}$$

Consequently, we find using (5.11) and (5.12) and by elementary manipulations that $D_+{}'(0)$ breaks into

$$D_+{}'(0) = \left(\frac{\partial(\Psi_{p+} \cdot \Phi_{p-})}{\partial \lambda} \det(\Psi_{i+} \cdot \Phi_{j-})_{1 \le i,j \le p-1} \det(\Psi_{i+} \cdot \Phi_{j-})_{p+1 \le i,j \le n} \right)_{|\lambda=0}. \tag{5.13}$$

The last determinant in (5.13) cannot be evaluated explicitly since the Ψ_{i+}, $i \in \{p+1, \dots, n\}$, are only known to vanish at $+\infty$, whereas the Φ_{j-}, $j \in \{p+1, \dots, n\}$, are known to vanish at $-\infty$. It is actually the counterpart of the transversality coefficient γ that appears in the derivative of the standard Evans function $D'(0)$ (see (6.1) below). On the other hand, the first two terms in (5.13) are related to the Majda–Liu determinant

$$M = (r_1^- \wedge \cdots \wedge r_{p-1}^- \wedge [u] \wedge r_{p+1}^+ \wedge \cdots \wedge r_n^+).$$

As a matter of fact, from the definition (5.10) of the Ψ_{i+}, we have

$$\begin{aligned} \det(\Psi_{i+} \cdot \Phi_{j-})_{1 \le i,j \le p-1} &= \det(-h_i A_- r_j^-)_{1 \le i,j \le p-1} \\ &= (-a_1^-) \cdots (-a_{p-1}^-) \det(h_i r_j^-)_{1 \le i,j \le p-1}. \end{aligned}$$

It remains to compute $\partial(\Psi_{p+} \cdot \Phi_{p-})/\partial\lambda|_{\lambda=0}$. Since $\partial\Psi_{p+}/\partial\lambda|_{\lambda=0}$ is at most algebraically growing and Φ_{p-} is exponentially decaying at $+\infty$, we have

$$\frac{\partial}{\partial \lambda}(\Psi_{p+} \cdot \Phi_{p-})_{|\lambda=0} = \lim_{+\infty} \left(\Psi_{p+} \cdot \frac{\partial \Phi_{p-}}{\partial \lambda} \right)_{|\lambda=0}.$$

Now by a standard computation (see [12]) we have

$$\frac{\partial \Phi_{p-}}{\partial \lambda}|_{\lambda=0} = \begin{pmatrix} \dfrac{\partial \phi_{p-}}{\partial \lambda}|_{\lambda=0} \\ B^{-1}A\dfrac{\partial \phi_{p-}}{\partial \lambda}|_{\lambda=0} + B^{-1}(U-u_-) \end{pmatrix}.$$

Hence, in view of (5.10) we have

$$\lim_{+\infty}\left(\Psi_{p+}\cdot\frac{\partial \Phi_{p-}}{\partial \lambda}\right)|_{\lambda=0} = h_p\,[u].$$

Finally, we obtain

$$\begin{aligned}&\left(\frac{\partial(\Psi_{p+}\cdot\Phi_{p-})}{\partial\lambda}\ \det(\Psi_{i+}\cdot\Phi_{j-})_{1\le i,j\le p-1}\right)_{|\lambda=0}\\ &\qquad = h_p\,[u]\ (-a_1^-)\cdots(-a_{p-1}^-)\ \det(h_i r_j^-)_{1\le i,j\le p-1}.\end{aligned}$$

This is related to M through the matrix identity

$$\begin{pmatrix} h_1 \\ \vdots \\ h_p \\ \ell_{p+1}^+ \\ \vdots \\ \ell_n^+ \end{pmatrix}\begin{pmatrix} r_1^-,\dots,r_{p-1}^-,[u],r_{p+1}^+,\dots,r_n^+ \end{pmatrix}$$

$$= \begin{pmatrix} h_1r_1^- & \cdots & h_1r_{p-1}^- & h_1[u] & \\ \vdots & \ddots & \vdots & \vdots & 0 \\ h_{p-1}r_1^- & \cdots & h_{p-1}r_{p-1}^- & h_{p-1}[u] & \\ 0 & \cdots & 0 & h_p[u] & \\ & & * & & I_{n-p} \end{pmatrix}.$$

Indeed we have

$$h_p\,[u]\ \det(h_i r_j^-)_{1\le i,j\le p-1} = (h_1\wedge\cdots\wedge h_p\wedge\ell_{p+1}^+\wedge\cdots\wedge\ell_n^+)\ M.$$

5.2. The extreme shock case. From now on, we consider an n-shock, that is, $p=n$. Of course, the case of 1-shock can be treated symmetrically by means of the mixed function D_- instead of D_+. Then the "undetermined determinant" in (5.13) does not appear. In view of Proposition 5.2 with $p=n$, we may choose

$$h_i = \ell_i^-,\quad i\in\{1,\dots,n\}.$$

Consequently, (5.13) reduces to

$$D_+{}'(0) = \ell_n^-\,[u]\ (-a_1^-)\cdots(-a_{n-1}^-),$$

if we normalize the eigenvectors by $\ell_i r_j = \delta_{ij}$. In particular, since the a_j^-, $j\in\{1,\dots,n-1\}$, are negative, we have the very simple result

$$\operatorname{sgn} D_+{}'(0) = \operatorname{sgn}\ \ell_n^-\,[u]. \tag{5.14}$$

Now let us determine the sign of $D_+(\lambda)$ for large real λ. We adapt the homotopy method developed in section 4.2, using the operator

$$L_\theta := \theta L + (1-\theta)\frac{d^2}{dx^2}, \quad \theta \in [0,1].$$

For large real λ and $\theta = 0$, i.e., when the viscosity matrix is the identity, the unstable manifold of

$$W' = \mathbb{A}(x;\lambda,0)W = \begin{pmatrix} 0 & I_n \\ \lambda & 0 \end{pmatrix} W$$

is spanned by

$$\Phi_{j-}(x;\lambda,0) = \begin{pmatrix} e^{x\sqrt{\lambda}}p_j^- \\ \sqrt{\lambda}e^{x\sqrt{\lambda}}p_j^- \end{pmatrix}$$

for some vectors p_j^-. Similarly, the stable manifold of

$$Z' = -Z\mathbb{A}(x;\lambda,0)$$

is spanned by

$$\Psi_{i+}(x;\lambda,0) = \left(\sqrt{\lambda}e^{-x\sqrt{\lambda}}q_i^+, e^{-x\sqrt{\lambda}}q_i^+\right)$$

for some co-vectors q_i^+. This implies that

$$D_+(\lambda,0) = (2\sqrt{\lambda})^n \det(q_i^+\, p_j^-) = (2\sqrt{\lambda})^n (q_1^+ \wedge \cdots \wedge q_n^+)(p_1^- \wedge \cdots \wedge p_n^-).$$

On one hand, we know from Lemma 7.3 in section 7 and a continuity argument that

$$\operatorname{sgn}\,(p_1^- \wedge \cdots \wedge p_n^-) = \operatorname{sgn}\,(r_1^- \wedge \cdots \wedge r_{n-1}^- \wedge s_n^-). \tag{5.15}$$

As a matter of fact, we recall that, for $\theta = 1$ and $\lambda = 0$, the unstable manifold of $\mathbb{A}_-$ is spanned by $(r_j^-, 0)^T$, $j \in \{1,\dots,n-1\}$, and $(s_n^-, \gamma_n^- s_n^-)^T$, where s_n^- is a right eigenvector of $B_-^{-1}A_-$ associated to γ_n^-. On the other hand, we have the following dual version of Lemma 7.2.

LEMMA 5.3. *Assuming* (H1) *and* (H5), *for all* $\theta \in [0,1]$, *for* $\lambda \in [0,+\infty[$, *the projection* $(\zeta, q) \mapsto q$ *is one-to-one from* $U_+^*(\lambda,\theta)$ *to* $\mathbb{R}^n$.

Note that for $\lambda = 0$, this can be viewed as a consequence of Proposition 5.2. Therefore, by Lemma 5.3 and a continuity argument, we have

$$\operatorname{sgn}\,(q_1^+ \wedge \cdots \wedge q_n^+) = \operatorname{sgn}\,(\ell_1^- B \wedge \cdots \wedge \ell_n^- B). \tag{5.16}$$

Since $\det B > 0$, we deduce from (5.15) and (5.16) that

$$\operatorname{sgn}\, D_+(\lambda,0) = \operatorname{sgn}\,(\ell_1^- \wedge \cdots \wedge \ell_n^-)(r_1^- \wedge \cdots \wedge r_{n-1}^- \wedge s_n^-),$$

and thus also by continuity

$$\operatorname{sgn}\, D_+(\lambda) = \operatorname{sgn}\,(\ell_1^- \wedge \cdots \wedge \ell_n^-)(r_1^- \wedge \cdots \wedge r_{n-1}^- \wedge s_n^-) \tag{5.17}$$

for large real λ. Finally, in view of (5.14) and (5.17), we have

$$\operatorname{sgn}\, D'_+(0)\; D_+(+\infty) = \operatorname{sgn}\,(r_1^- \wedge \cdots \wedge r_{n-1}^- \wedge s_n^-)(\ell_1^- \wedge \cdots \wedge \ell_n^-)\; \ell_n^-\,[u].$$

It is easy to check, by using the relations $\ell_i r_j = \delta_{ij}$, that this sign is the same as the standard one (see (4.16) in the case $n = 2$, and (7.7) below in the general case).

6. Cases of neutral instability. In this section, we focus on the case of a Lax shock for which $D'(0)$ vanishes. This is a borderline case, where neighboring shocks may have an odd or an even index.

6.1. An interesting consequence of neutral instability. Let us first recall the mechanism for this "neutral" instability to occur. (We refer to [25] for a multidimensional version.) We use here the standard Evans function (2.2), which equivalently reads

$$D(\lambda) = e^{-\int_0^x \operatorname{tr}\mathbb{A}(s;\lambda)ds}\,(\Phi_{1-}(x;\lambda)\wedge\cdots\wedge\Phi_{n-}(x;\lambda)\wedge\Phi_{1+}(x;\lambda)\wedge\cdots\wedge\Phi_{n+}(x;\lambda)).$$

We recall that the Φ_j are solutions of the variable coefficient ODE (2.3), where

$$\mathbb{A}(\cdot;\lambda) = \begin{pmatrix} 0_n & I_n \\ B^{-1}(\lambda + A') & B^{-1}(A - B') \end{pmatrix},$$

with $A(x)$ and $B(x)$ defined by (3.4) and (3.5), respectively. We have the following lemma.

LEMMA 6.1. *Assuming* (H1) *and* (H5) *and that the discontinuity* (u_-, u_+) *is a p-shock, that is,* (5.1) *and* (5.2) *hold, we can choose the* Φ_j *such that, for* $\lambda = 0$,

$$\begin{aligned} \Phi_{p-} = \Phi_{p+} &= \begin{pmatrix} U' \\ U'' \end{pmatrix}, \\ \Phi_{j-}(-\infty) &= \begin{pmatrix} r_j^- \\ 0 \end{pmatrix}, \quad j \in \{1,\dots,p-1\}, \\ \Phi_{j+}(+\infty) &= \begin{pmatrix} r_j^+ \\ 0 \end{pmatrix}, \quad j \in \{p+1,\dots,n\}, \end{aligned}$$

where the $r_j^\pm$ *are right eigenvectors of* $A_\pm$ *associated with* $a_j^\pm$. *Then we have*

$$D'(0) = (-1)^n \det B(0)^{-1}(-a_1^-)\cdots(-a_{p-1}^-)(a_{p+1}^+)\cdots(a_n^+)\,\gamma\, M, \tag{6.1}$$

where M *is the Majda–Liu determinant*

$$M := (r_1^- \wedge\cdots\wedge r_{p-1}^- \wedge [u] \wedge r_{p+1}^+ \wedge\cdots\wedge r_n^+) \tag{6.2}$$

and γ *is given by*

$$\gamma := e^{-\int_0^x \operatorname{tr}(B^{-1}A)}\,(\phi_{1+}(x;0)\wedge\cdots\wedge\phi_{(p-1)+}(x;0)\wedge\phi_{p-}(x;0)\wedge\cdots\wedge\phi_{n-}(x;0)). \tag{6.3}$$

Furthermore, provided that the family $\{a_1^- r_1^-,\dots,a_{p-1}^- r_{p-1}^-, a_{p+1}^+ r_{p+1}^+,\dots,a_n^+ r_n^+\}$ *is independent, we have* $D'(0) = 0$ *if and only if there exist coefficients* $b_j^\pm$ *such that*

$$[u] = \sum_{j=1}^{p-1} b_j^-\, a_j^- r_j^- + \sum_{j=p+1}^{n} b_j^+\, a_j^+ r_j^+, \tag{6.4}$$

and, denoting $\psi_{p\pm} = \partial\phi_{p\pm}/\partial\lambda$,

$$\psi_{p+} - \psi_{p-} = b_p\, U' + \sum_{j\neq p} b_j^-\,\phi_{j-} + \sum_{j\neq p} b_j^+\,\phi_{j+} \tag{6.5}$$

at $\lambda = 0$.

Proof. The special choice of Φ_j is similar as in section 5.1 (see also [12]). Let us briefly recall the computation of $D'(0)$. Since $\Phi_{p-} = \Phi_{p+}$, we have just

$$\begin{aligned} D'(0) = e^{-\int_0^x \operatorname{tr}\mathbb{A}(s;0)ds}(&\Phi_{1-}(x;0)\wedge\cdots\wedge\Phi_{n-}(x;0)\wedge\Phi_{1+}(x;0)\wedge\cdots\wedge\Phi_{(p-1)+}(x;0)\\ &\wedge\Psi_{p+}(x;0) - \Psi_{p-}(x;0)\wedge\Phi_{(p+1)+}(x;0)\wedge\cdots\wedge\Phi_{n+}(x;0)), \end{aligned} \tag{6.6}$$

where $\Psi_{p\pm} := \partial\Phi_{p\pm}/\partial\lambda$. By construction, $\phi_{j+}(\cdot;0)$, $j \in \{1,\dots,p\}$, as well as the $\phi_{j-}(\cdot;0)$, $j \in \{p,\dots,n\}$, are solutions of the variable coefficient ODE

$$\phi' = B^{-1}A\phi.$$

On the other hand, we have

$$B\,\phi'_{j\pm} - A\,\phi_{j\pm} \equiv -a_j^{\pm}\,r_j^{\pm}$$

for $j \in \{1,\dots,p-1\}$ (with the sign $-$) or $j \in \{p+1,\dots,n\}$ (with the sign $+$) at $\lambda = 0$, and

$$B\,\psi'_{p\pm} - A\,\psi_{p\pm} = U - u_{\pm}.$$

Using these relations we obtain a block-triangular matrix by multiplying the matrix in (6.6) by

$$\begin{pmatrix} I_n & 0_n \\ -A(x) & B(x) \end{pmatrix}.$$

More precisely, using the remark that

$$e^{-\int_0^x \operatorname{tr}\mathbb{A}(s;0)ds} = e^{-\int_0^x \operatorname{tr}(B^{-1}(A-B'))} = \frac{\det B(x)}{\det B(0)}\, e^{-\int_0^x \operatorname{tr}(B^{-1}A)}$$

and permuting $(p-1)$ columns we obtain

$$D'(0) = \frac{e^{-\int_0^x \operatorname{tr}(B^{-1}A)}}{\det B(0)}\,(-1)^{p-1}\,\det\begin{pmatrix} \Gamma & * \\ 0_n & \Delta \end{pmatrix},$$

where

$$\Gamma := \left(\phi_{1+}(x;0),\dots,\phi_{(p-1)+}(x;0),\phi_{p-}(x;0),\dots,\phi_{n-}(x;0)\right),$$

and

$$\Delta := \left(-a_1^- r_1^-,\dots,-a_{p-1}^- r_{p-1}^-,-[u],-a_{p+1}^+ r_{p+1}^+,\dots,-a_n^+ r_n^+\right).$$

Then the result in (6.1) follows immediately, and the end of the proof comes from a pure algebraic argument. ∎

Note that if the eigenvalues $a_j^{\pm}$ are nonzero, the independence of the family $\{a_1^- r_1^-,\dots,a_{p-1}^- r_{p-1}^-, a_{p+1}^+ r_{p+1}^+,\dots,a_n^+ r_n^+\}$ is ensured by the transversality condition (5.7). Also note that γ itself is a Wronskian. Actually, γ measures transversality of the stable/unstable manifolds of the traveling wave ODE at u_+/u_-.

Now, let us examine the situation when we have such decompositions as (6.4) and (6.5). If the eigenvalues $a_j^{\pm}$ are nonzero and the transversality condition (5.7) holds,

(6.4) implies that the Majda–Liu determinant M is equal to 0. Shocks that satisfy this property cannot be of small amplitude, at least in regions where the system is strictly hyperbolic. However, weak shocks of this kind occur in MHD near states with crossing eigenvalues. When zero is a regular value of M, viewed as a function of the right state u_+, for instance, its zeros form a hypersurface which may be viewed as a transition locus between multidimensionally stable and unstable shock waves (see [22]). Here stability refers to the hyperbolic system $u_t + \operatorname{div} F = 0$, with $F_1 = f$.

Assuming that a given shock satisfies $M = 0$ and that it admits a profile U, we have $D'(0) = 0$ and we still know how to calculate the sign of $D(\lambda)$ for large real λ (see section 7). Therefore, we face the calculation of $D''(0)$. If $D''(0)D(+\infty) < 0$, we shall be able to conclude to the linear instability of the profile (more precisely, the "strong instability" of the profile: the double root of $D(\cdot)$ at $\lambda = 0$ already provides a "neutral instability"; see [12]). An interesting consequence will then happen: the linear operator admits a real positive eigenvalue λ_0, where the sign of D changes (because of analyticity, since one of the real roots of D must have an odd multiplicity). Let us move the right state u_+ slightly along the Hugoniot curve of u_-, denoted by $H_p(u_-)$. When z moves along $H_p(u_-)$, in a neighborhood of u_+, the above determinant vanishes only if $z = u_+$ and its sign changes across u_-. Therefore, we are free to choose z in such a way that, for the perturbed shock (u_-, z), the sign of $D'(0; z)D(+\infty; z)$ is positive. Then the number of real positive eigenvalues of $D(\cdot; z)$ is even, counting with multiplicity, but the eigenvalue λ_0 persists, with a small perturbation, if z is close enough to u_+. This construction provides examples where we are able to conclude to linear instability, although the index of the linearized operator is even.

6.2. Computation of $D''(0)$ when $D'(0) = 0$. We propose a method for computing $D''(0)$ when we have (6.4) and (6.5). The outcome will be an integral formula, which is (at least numerically) plainly computable.

A useful trick consists in noting that D is unchanged by adding to $\Phi_{p\pm}$ a linear combination of the other Φs. More precisely, defining

$$\tilde{\Phi}_{p+} := \Phi_{p+} + \lambda \left(b_p \, \Phi_{p+} + \sum_{j \neq p} b_j^+ \, \Phi_{j+} \right),$$

$$\tilde{\Phi}_{p-} := \Phi_{p-} - \lambda \left(\sum_{j \neq p} b_j^- \, \Phi_{j-} \right),$$

we have

$$D(\lambda) = (\Phi_{1-} \wedge \cdots \wedge \tilde{\Phi}_{p-} \wedge \cdots \wedge \Phi_{n-} \wedge \Phi_{1+} \wedge \cdots \wedge \tilde{\Phi}_{p+} \wedge \cdots \wedge \Phi_{n+})|_{x=0}.$$

Denoting naturally $\tilde{\psi}_{p\pm} := \partial \tilde{\phi}_{p\pm}/\partial\lambda$, we have because of (6.5), $\tilde{\psi}_{p+} = \tilde{\psi}_{p-}$ at $\lambda = 0$. This is the function that will appear in the final integral formula. We simply denote it by $\tilde{\Psi}$. Now, denoting

$$\tilde{\zeta}_{p\pm} := \frac{1}{2} \left. \frac{\partial^2 \tilde{\phi}_{p\pm}}{\partial \lambda^2} \right|_{\lambda=0},$$

we have

$$\left(B \, \tilde{\zeta}'_{p\pm} - A \, \tilde{\zeta}_{p\pm} \right)' = \tilde{\psi}. \tag{6.7}$$

This implies in particular that

$$B\,(\tilde\zeta'_{p+}-\tilde\zeta'_{p-})-A\,(\tilde\zeta_{p+}-\tilde\zeta_{p-})\equiv C, \tag{6.8}$$

a constant. By similar operations as in the computation of $D'(0)$, thus we find that

$$\frac12\,D''(0)=(-1)^{n-1}\,\det B(0)^{-1}(-a_1^-)\cdots(-a_{p-1}^-)(a_{p+1}^+)\cdots(a_n^+)\,\gamma\,I, \tag{6.9}$$

where

$$I:=(r_1^-\wedge\cdots\wedge r_{p-1}^-\wedge C\wedge r_{p+1}^+\wedge\cdots\wedge r_n^+). \tag{6.10}$$

The problem in evaluating I is that we do not have access to the constant C. However, we shall see that I can be written as an integral in terms of $\tilde\psi$, plus boundary terms that are explicit at least in the case of an extreme shock ($p=1$ or $p=n$).

Assuming the transversality condition (5.7), we can preferably write $I=\ell\cdot C$, where ℓ is a co-vector such that

$$\ell\cdot r_1^-=0,\dots,\ell\cdot r_{p-1}^-=0,\quad \ell\cdot r_{p+1}^+=0,\dots,\ell\cdot r_n^+=0. \tag{6.11}$$

With the notations of section 5.1, we have

$$\ell=h_p\,(r_1^-\wedge\cdots\wedge r_{p-1}^-\wedge r_p^\pm\wedge r_{p+1}^+\wedge\cdots\wedge r_n^+)/(h_p\cdot r_p^\pm).$$

Then (6.7) and (6.8) show that

$$\begin{aligned} I=\ell\cdot C=&-\int_{-\infty}^{+\infty}\ell\cdot\tilde\psi(x)\,dx\\ &+\Big(\ell\cdot(B\,\tilde\zeta'_{p+}-A\,\tilde\zeta_{p+})\Big)(+\infty)-\Big(\ell\cdot(B\,\tilde\zeta'_{p-}-A\,\tilde\zeta_{p-})\Big)(-\infty). \end{aligned} \tag{6.12}$$

Let us comment on this formula. The first part is similar to a formula obtained by Kapitula in [16]. However, there are no boundary terms in Kapitula's formula. Here they come from the fact that 0 is embedded in the essential spectrum (by the translational invariance). The integral in (6.12) is computable provided that $\tilde\psi$ is, which is at least possible numerically. Indeed, $\tilde\psi$ is the solution of the "boundary value problem"

$$(B\tilde\psi'-A\,\tilde\psi)'=U',\quad \tilde\psi(-\infty)=-\sum_{j=1}^{p-1}b_j^-\,r_j^-,\quad \tilde\psi(+\infty)=\sum_{j=p+1}^{n}b_j^+\,r_j^+.$$

Note that these boundary conditions and equations in (6.11) ensure that $\ell\cdot\tilde\psi$ is integrable (whereas $\tilde\psi$ is not!). Concerning the boundary terms in (6.12), we can give more explicit formulas. As a matter of fact, we have

$$\tilde\zeta_{p+}=\zeta_{p+}+b_p\;\psi_{p+}+\sum_{j\neq p}b_j^+\,\psi_{j+},\quad \tilde\zeta_{p-}=\zeta_{p-}-\sum_{j\neq p}b_j^-\,\psi_{j-},$$

where

$$\zeta_{p\pm}:=\frac12\,\frac{\partial^2\tilde\phi_{p\pm}}{\partial\lambda^2}\bigg|_{\lambda=0},\quad \psi_{j\pm}:=\frac{\partial\phi_{j\pm}}{\partial\lambda}\bigg|_{\lambda=0}.$$

Hence, we have

$$\begin{aligned}\tilde{\zeta}_{p+}(x) &\sim \sum_{j=p+1}^{n} b_j^+\,\psi_{j+}(x), \quad x \to +\infty \quad (\text{if } p \le n-1),\\ \tilde{\zeta}_{p-} &\sim -\sum_{j=1}^{p-1} b_j^-\,\psi_{j-}(x), \quad x \to -\infty \quad (\text{if } p \ge 2).\end{aligned} \tag{6.13}$$

Thus the boundary terms in (6.12) depend on the asymptotic behaviors of $\psi_{j\pm}$. Recalling that

$$\phi_{j\pm}(x;\lambda) \sim e^{\nu_j^\pm(\lambda)x}\, v_j^\pm(\lambda), \quad x \to \pm\infty,$$

with $(\nu_j^\pm(\lambda), v_j^\pm(\lambda))$ such that

$$\nu_j^\pm(\lambda) = -\frac{\lambda}{a_j^\pm} + \mathcal{O}(\lambda^2), \quad v_j^\pm(\lambda) = r_j^\pm$$

for $j \in \{1,\dots,p-1\}$ (with the sign $-$) or $j \in \{p+1,\dots,n\}$ (with the sign $+$), we have

$$\psi_{j\pm}(x) \sim -\frac{x}{a_j^\pm}\, r_j^\pm + \left.\frac{dv_j^\pm}{d\lambda}\right|_{\lambda=0}, \quad x \to \pm\infty.$$

Therefore, we have using (6.11)

$$\left(\ell \cdot (B\,\psi'_{j\pm} - A\,\psi_{j\pm})\right)(\pm\infty) = \ell \cdot \left(-\frac{1}{a_j^\pm}\, B_\pm\, r_j^\pm - A_\pm \left.\frac{dv_j^\pm}{d\lambda}\right|_{\lambda=0}\right) \tag{6.14}$$

for $j \in \{1,\dots,p-1\}$ (with the sign $-$) or $j \in \{p+1,\dots,n\}$ (with the sign $+$). Using (6.13) and (6.14) we obtain by linearity the value of the boundary terms in (6.12). It is not very simple but it is "explicit." Indeed, the vectors

$$\left.\frac{dv_j^\pm}{d\lambda}\right|_{\lambda=0}$$

can be computed by differentiating twice the relation satisfied by $(\nu_j^-(\lambda), v_j^-(\lambda))$,

$$(\nu^2\, B_\pm - \nu\, A_\pm - \lambda)\, v = 0.$$

Actually, the formula is much simpler for extreme shocks. Assume, for instance, that $p = n$. Then there is only the boundary term at $-\infty$ in (6.12). Furthermore, ℓ is necessarily colinear to ℓ_n^-. This implies by the double differentiation mentioned above that

$$\ell \cdot \left(-\frac{1}{a_j^-}\, B_-\, r_j^-\right) = (a_n^- - a_j^-)\, \ell \cdot \left.\frac{dv_j^\pm}{d\lambda}\right|_{\lambda=0}$$

for all $j \in \{1,\dots,n-1\}$. Substituting this equality in (6.14) we find that

$$\left(\ell \cdot (B\,\psi'_{j-} - A\,\psi_{j-})\right)(-\infty) = \frac{(\ell \cdot B_-\, r_j^-)}{a_n^- - a_j^-}.$$

Finally, we can summarize our result in the following theorem.

THEOREM 6.2. *We assume that* (H1) *and* (H5) *hold true and that the discontinuity* (u_-, u_+) *is a* n*-shock, that is,* (5.1) *and* (5.2) *hold with* $p = n$*. We also assume that the family* $\{a_1^- r_1^-, \dots, a_{n-1}^- r_{n-1}^-\}$ *is independent and that the functions* $\Phi_j^\pm$ *are chosen as in Lemma* 6.1*. If* $D'(0) = 0$*, then there exist a unique family* $\{b_1, \dots, b_n\}$ *such that*

$$[u] = \sum_{j=1}^{n-1} b_j\, a_j^- r_j^-. \tag{6.15}$$

Furthermore, we have

$$\frac{1}{2}\, D''(0) = (-1)^{n-1}\, \det B(0)^{-1} (-a_1^-) \cdots (-a_{p-1}^-)(a_{p+1}^+) \cdots (a_n^+)\, \gamma\, I, \tag{6.16}$$

where γ *is defined by* (6.3) *and*

$$\begin{aligned} I = \; & -\int_{-\infty}^{+\infty} (r_1^- \wedge \cdots \wedge r_{n-1}^- \wedge \tilde\psi(x))\, dx \\ & + \left(r_1^- \wedge \cdots \wedge r_{n-1}^- \wedge \sum_{j=1}^{n-1} \frac{b_j}{a_n^- - a_j^-}\, B_-\, r_j^- \right) \end{aligned} \tag{6.17}$$

with $\tilde\psi$ *a function such that*

$$(B\tilde\psi' - A\,\tilde\psi)' = U', \quad \tilde\psi(-\infty) = -\sum_{j=1}^{n-1} b_j\, r_j^-,\ \tilde\psi(+\infty) = 0.$$

Note that $(U', \tilde\psi)$ form a Jordan chain for the operator L, since we have

$$L \cdot U' = 0, \quad L \cdot \tilde\psi = U'.$$

7. The case of $n \times n$ systems. In this section, we are interested in the question left open by Gardner and Zumbrun (see [12, p. 837]), concerning $n \times n$ systems. More precisely, there was missing an n-dimensional version of their "algebraic" Lemma 3.5, needed to complete the theory for $n > 2$. We shall prove such a lemma under the following hypotheses.

(h1) The $n \times n$ matrix A is real-valued, invertible, and there exists a positive definite symmetric matrix S_0 such that $S_0 A$ is symmetric;

(h2) the $n \times n$ matrix B is real-valued and there exists $\beta > 0$ such that

$$\text{for all } X \in \mathbb{C}^n, \quad \text{Re}\, \langle BX, X\rangle \geq \beta\, \langle X, X\rangle\,, \tag{7.1}$$

where $\langle ., . \rangle$ is the inner product associated to S_0.

In particular, (h1) and (h2) hold if $A = df(u)$, $B = B(u)$, and (H5) of section 3 holds in the neighborhood of the constant state u. Let us recall from Remark 2 that this is the case for most physical systems of conservation laws.

We shall denote by $\mathcal{U}(A)$ the unstable manifold of A and by $\mathcal{S}(B^{-1}A)$ the stable manifold of $B^{-1}A$. These are a priori subspaces of $\mathbb{C}^n$. However, since A and B are real-valued, we may also consider them as subspaces of $\mathbb{R}^n$. Let us first show the following lemma.

LEMMA 7.1. *Assuming* (h1) *and* (h2), *the subspaces* $\mathcal{U}(A)$ *and* $\mathcal{S}(B^{-1}A)$ *are complementary.*

Proof. The assumptions (h1) and (h2) imply that A, B satisfy the properties required in (H2), (H3), and (H4). Moreover, A satisfies the weakened assumption (H1′). Hence, by Lemma 3.2 and Remark 3, we have

$$n = \dim \mathcal{U}(A) + \dim \mathcal{S}(B^{-1}A).$$

It remains to show that these subspaces have their intersection reduced to 0. Let $X_0 \neq 0$ belong to the stable manifold of $B^{-1}A$. By definition, there exists φ such that

$$\varphi' = B^{-1}A\varphi$$

and

$$\begin{cases} \varphi(+\infty) = 0, \\ \varphi(0) = X_0. \end{cases}$$

By the symmetry of A, we have

$$\langle A\varphi, \varphi\rangle' = 2\mathrm{Re}\,\langle A\varphi, B^{-1}A\varphi\rangle \geq 2\beta|B^{-1}A\varphi|^2.$$

Since $X_0 \neq 0$ and A is invertible, we conclude that $\langle A\varphi, \varphi\rangle < 0$ on $[0, +\infty[$. In particular,

$$\langle AX_0, X_0\rangle < 0.$$

On the other hand, if X belongs to the unstable manifold of A, which is diagonalizable in an orthogonal basis, then we must have $\langle AX, X\rangle \geq 0$. ∎

For $\lambda \in \mathbb{C}$, we take the usual notation

$$\mathbb{A}(\lambda) = \begin{pmatrix} 0_n & I_n \\ \lambda B^{-1} & B^{-1}A \end{pmatrix}.$$

For $\mathrm{Re}\,\lambda > 0$, we denote by $S(\lambda)$, respectively, $U(\lambda)$, the stable/unstable manifold of $\mathbb{A}(\lambda)$. They are then extended by continuity to $\lambda = 0$.

LEMMA 7.2. *Assuming* (h1) *and* (h2), *for all* $\lambda \in [0, +\infty[$, *the projection* $(v, w)^T \mapsto v$ *is one-to-one from* $S(\lambda)$ *to* $\mathbb{R}^n$.

Proof. Since (h1) and (h2) imply the Majda–Pego condition (H4), we know that $\mathbb{A}(\lambda)$ does not have a center manifold for $\mathrm{Re}\,\lambda > 0$. The far field behavior and a continuity argument then show that for $\mathrm{Re}\,\lambda > 0$

$$\dim S(\lambda) = \dim U(\lambda) = n.$$

By continuity, it also holds for $\lambda = 0$. Thus it is sufficient to show that the projection $(v, w)^T \mapsto v$ has zero kernel in $S(\lambda)$ for $\mathrm{Re}\,\lambda > 0$ or $\lambda = 0$.

First let us treat the case $\lambda = 0$. It is known from standard computations (see [12]) that $S(0)$ consists of a center part and a genuine stable part. The stable part is spanned by the generalized eigenvectors associated to the eigenvalues of negative real part of $\mathbb{A}(0)$, which coincide with the eigenvalues of negative real part of $B^{-1}A$. The v-component as well as the w-component of these generalized eigenvectors belong to $\mathcal{S}(B^{-1}A)$. The center part is derived through a bifurcation analysis. It is spanned by

the vectors $(r,0)^T$, with $r \in \mathcal{U}(A)$. Let us consider any vector $(v_s + r, w_s)^T \in S(0)$, with $v_s, w_s \in \mathcal{S}(B^{-1}A)$ and $r \in \mathcal{U}(A)$. If $v_s + r = 0$, then by Lemma 7.1, we have necessarily $v_s = 0$ and $r = 0$. Since $(v_s = 0, w_s)^T$ belongs to the stable manifold of $\mathbb{A}(0)$, there should exist (ϕ, φ) such that

$$\begin{cases} \phi' = \varphi, & \varphi' = B^{-1}A\varphi, \\ \phi(0) = 0, & \varphi(0) = w_s, \\ \phi(+\infty) = 0, & \varphi(+\infty) = 0. \end{cases}$$

By integration, this implies $w_s = 0$. This proves the lemma in the case $\lambda = 0$.

We now assume $\operatorname{Re}\lambda > 0$. We shall use a *Lyapunov function* to show the following preliminary result.

- If $(v,0)^T \in S(\lambda)$, then $v = 0$.

As a matter of fact, let us define

$$H(v,w) := \frac{1}{2}\langle Aw, w\rangle + \operatorname{Re}\langle w, \lambda v\rangle.$$

If one has

$$\begin{pmatrix} \phi' \\ \varphi' \end{pmatrix} = \mathbb{A}(\lambda)\begin{pmatrix} \phi \\ \varphi \end{pmatrix}, \tag{7.2}$$

then by the symmetry of A

$$H(\phi,\varphi)' = \operatorname{Re}\left\langle \lambda\phi + A\varphi, B^{-1}(\lambda\phi + A\varphi)\right\rangle + \operatorname{Re}\lambda|\varphi|^2.$$

Hence, by (h2), we have $H(\phi,\varphi)' \geq 0$ with equality if and only if

$$\begin{cases} \lambda\phi + A\varphi = 0, \\ \operatorname{Re}\lambda\,|\varphi| = 0, \end{cases}$$

which is equivalent to $(\phi,\varphi) = (0,0)$. Suppose that $(v_0,0)^T \in S(\lambda)$. Then there exists (ϕ,φ) solution to (7.2) such that

$$\begin{cases} \phi(0) = v_0, & \varphi(0) = 0, \\ \phi(+\infty) = 0, & \varphi(+\infty) = 0. \end{cases}$$

Since

$$H(v_0,0) = H(0,0) = 0,$$

and H is nondecreasing along solutions of (7.2), we must have $H(\phi,\varphi) \equiv 0$ on $[0,+\infty[$. Thus we also have $H(\phi,\varphi)' \equiv 0$ on $[0,+\infty[$, which implies that $(\phi,\varphi) \equiv (0,0)$ on $[0,+\infty[$. In particular, this shows that $v_0 = 0$.

We are now in a position to show that

- if $(0,w)^T \in S(\lambda)$, then $w = 0$.

Since $S(\lambda)$ is invariant under $\mathbb{A}(\lambda)$, which is invertible, if $(0,w)^T \in S(\lambda)$, there exists $(v_1,w_1)^T \in S(\lambda)$ such that

$$\begin{pmatrix} 0 \\ w \end{pmatrix} = \mathbb{A}(\lambda)\begin{pmatrix} v_1 \\ w_1 \end{pmatrix}.$$

We actually have $0 = w_1$. From the previous result, this implies $v_1 = 0$, and thus also $w = 0$. This concludes the proof. ∎

Of course, we have symmetrically the following lemma.

LEMMA 7.3. *Assuming* (h1) *and* (h2), *for all* $\lambda \in [0, +\infty[$, *the projection* $(v, w)^T \mapsto v$ *is one-to-one from* $U(\lambda)$ *to* $\mathbb{R}^n$.

Lemmas 7.2 and 7.3 prove, in particular, that the stability conditions stated in [12, p. 837], do hold. More generally, we have the following theorem.

THEOREM 7.4. *Assuming* (H1) *and* (H5), *a necessary condition for linearized stability of a viscous n-shock wave* U *is*

$$(r_1^- \wedge \cdots \wedge r_{n-1}^- \wedge s_n^-)\ (r_1^- \wedge \cdots \wedge r_{n-1}^- \wedge [u]) \geq 0,$$

where s_n^- *denotes the asymptotic direction of* U' *at* $-\infty$.

Proof. We use the standard Evans function (2.2) and apply Lemma 6.1 regarding the long wave analysis. In the case $p = n$, γ reduces to

$$\gamma = e^{-\int_0^x \operatorname{tr}(B^{-1}A)}\ (\phi_{1+}(x;0) \wedge \cdots \wedge \phi_{n+}(x;0)).$$

Since it is a Wronskian, and the $\Phi_{j+}(\cdot;0)$ are asymptotic to the stable manifold, $S_+(0)$, of $\mathbb{A}_+(0)$ at $+\infty$, we may evaluate sgn γ at $+\infty$. Therefore, we have

$$\operatorname{sgn}\ \gamma = \operatorname{sgn}\ (s_1^+ \wedge \cdots \wedge s_n^+),$$

where the s_j^+ span $\mathcal{S}(B_+^{-1}A_+) = \mathbb{R}^n$. Since the (a_j^-), $j \in \{1, \ldots, n-1\}$, are negative, and

$$M = (r_1^- \wedge \cdots \wedge r_{n-1}^- \wedge [u]),$$

thus we infer that

$$\operatorname{sgn}\ D'(0) = \operatorname{sgn}\ (-1)^n\ (r_1^- \wedge \cdots \wedge r_{n-1}^- \wedge [u])\ (s_1^+ \wedge \cdots \wedge s_n^+). \tag{7.3}$$

On the other hand, applying either a homotopy (as in sections 4 and 5) or a rescaling method [12], we find that, for large real λ,

$$\operatorname{sgn}\ D(\lambda) = \operatorname{sgn}\ (-1)^n (v_1^- \wedge \cdots \wedge v_n^-)\ (v_1^+ \wedge \cdots \wedge v_n^+). \tag{7.4}$$

Here the $v_j^\pm$ are the projections onto the v-components of the bases of $U_-(\lambda)$ and $S_+(\lambda)$ provided by the asymptotic behavior of the Φ_{j-} and of the Φ_{j+}, respectively. From Lemma 7.2, a continuity argument and the asymptotics of the $\Phi_{j+}(\cdot;0)$, we have

$$\operatorname{sgn}\ (v_1^+ \wedge \cdots \wedge v_n^+) = \operatorname{sgn}\ (s_1^+ \wedge \cdots \wedge s_n^+). \tag{7.5}$$

Similarly, we can use Lemma 7.3 to find sgn $(v_1^- \wedge \cdots \wedge v_n^-)$. We recall that

$$\Phi_{j-}(-\infty;0) = \begin{pmatrix} r_j^- \\ 0 \end{pmatrix}, \quad j \in \{1, \ldots, n-1\}.$$

Therefore, if s_n^- denotes the asymptotic direction of $U' = \phi_{n-}(\cdot;0)$ at $-\infty$, which spans $\mathcal{S}(B_-^{-1}A_-)$, we have

$$\operatorname{sgn}\ (v_1^- \wedge \cdots \wedge v_n^-) = \operatorname{sgn}\ (r_1^- \wedge \cdots \wedge r_{n-1}^- \wedge s_n^-). \tag{7.6}$$

In view of (7.3), (7.4), (7.5), and(7.6), we have

$$\operatorname{sgn}\ D'(0)\ D(+\infty) = (r_1^- \wedge \cdots \wedge r_{n-1}^- \wedge [u])\ (r_1^- \wedge \cdots \wedge r_{n-1}^- \wedge s_n^-). \tag{7.7}$$

This concludes the proof of Theorem 7.4 by using the intermediate value theorem. ∎

REFERENCES

[1] J.C. Alexander, R. Gardner, and C.K.R.T. Jones, *A topological invariant arising in the stability analysis of travelling waves*, J. Reine Angew. Math., 410 (1990), pp. 167–212.

[2] J.C. Alexander, M.G. Grillakis, C.K.R.T. Jones, and B. Sandstede, *Stability of pulses on optical fibers with phase-sensitive amplifiers*, Z. Angew. Math. Phys., 48 (1997), pp. 175–192.

[3] S. Benzoni-Gavage, *On the stability of semidiscrete shock profiles by means of an Evans function in infinite dimension*, C. R. Acad. Sci. Paris Sér. I Math., 329 (1999), pp. 377–382.

[4] T.J. Bridges and G. Derks, *Hodge duality and the Evans function*, Phys. Lett. A, 251 (1999), pp. 363–372.

[5] J.W. Evans, *Nerve axon equations.* I. *Linear approximations*, Indiana Univ. Math. J., 21 (1971/1972), pp. 877–885.

[6] J.W. Evans, *Nerve axon equations.* II. *Stability at rest*, Indiana Univ. Math. J., 22 (1972/1973), pp. 75–90.

[7] J.W. Evans, *Nerve axon equations.* III. *Stability of the nerve impulse*, Indiana Univ. Math. J., 22 (1972/1973), pp. 577–593.

[8] J.W. Evans, *Nerve axon equations.* IV. *The stable and the unstable impulse*, Indiana Univ. Math. J., 24 (1974/1975), pp. 1169–1190.

[9] H. Freistühler and K. Zumbrun, *Examples of Unstable Viscous Shock Waves*, Technical report, Institut für Mathematik, Rheinisch-Westfälische Technische Hochschule, Aachen, Germany, 1998.

[10] R.A. Gardner and C.K.R.T. Jones, *Traveling waves of a perturbed diffusion equation arising in a phase field model*, Indiana Univ. Math. J., 39 (1990), pp. 1197–1222.

[11] R.A. Gardner and C.K.R.T. Jones, *A stability index for steady state solutions of boundary value problems for parabolic systems*, J. Differential Equations, 91 (1991), pp. 181–203.

[12] R.A. Gardner and K. Zumbrun, *The gap lemma and geometric criteria for instability of viscous shock profiles*, Comm. Pure Appl. Math., 51 (1998), pp. 797–855.

[13] D. Henry, *Geometric Theory of Semilinear Parabolic Equations*, Springer-Verlag, Berlin, 1981.

[14] C.K.R.T. Jones, *Stability of the travelling wave solution of the FitzHugh-Nagumo system*, Trans. Amer. Math. Soc., 286 (1984), pp. 431–469.

[15] C.K.R.T. Jones, R. Gardner, and T. Kapitula, *Stability of travelling waves for nonconvex scalar viscous conservation laws*, Comm. Pure Appl. Math., 46 (1993), pp. 505–526.

[16] T. Kapitula, *The Evans function and generalized Melnikov integrals*, SIAM J. Math. Anal., 30 (1998), pp. 273–297.

[17] T. Kapitula and B. Sandstede, *Stability of bright solitary-wave solutions to perturbed nonlinear Schrödinger equations*, Phys. D, 124 (1998), pp. 58–103.

[18] S. Kawashima, *Systems of a Hyperbolic–Parabolic Composite Type, with Applications to the Equations of Magnetohydrodynamics*, Ph.D. thesis, Kyoto University, Kyoto, 1983.

[19] A. Majda and R.L. Pego, *Stable viscosity matrices for systems of conservation laws*, J. Differential Equations, 56 (1985), pp. 229–262.

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

[21] D. Serre, *Systèmes de lois de conservation.* I, in Hyperbolicité, entropies, ondes de choc, Diderot Editeur, Paris, 1996.

[22] D. Serre, *La transition vers l'instabilité pour les ondes de choc multi-dimensionnelles*, http://www.umpa.ens-lyon.fr/ serre/PS/sw.ps (1999).

[23] J. Swinton, *The stability of homoclinic pulses: A generalisation of Evans's method*, Phys. Lett. A, 163 (1992), pp. 57–62.

[24] K. Zumbrun and P. Howard, *Pointwise semigroup methods and stability of viscous shock waves*, Indiana Univ. Math. J., 47 (1998), pp. 741–871.

[25] K. Zumbrun and D. Serre, *Viscous and inviscid stability of multidimensional planar shock fronts*, Indiana Univ. Math. J., 48 (1999), pp. 937–992.

SIAM J. MATH. ANAL.
Vol. 32, No. 5, pp. 963–986

DETERMINING COEFFICIENTS IN A CLASS OF HEAT EQUATIONS VIA BOUNDARY MEASUREMENTS*

B. CANUTO† AND O. KAVIAN†

Abstract. When $\Omega \subset \mathbb{R}^N$ is a bounded domain, we consider the problem of identifiability of the coefficients ρ, A, q in the equation $\rho(x)\partial_t u - \operatorname{div}(A(x)\nabla u) + q(x)u = 0$ from boundary measurements on two pieces Γ_{in} and Γ_{out} of $\partial\Omega$. Provided that $\Gamma_{\text{in}} \cap \Gamma_{\text{out}}$ has a nonempty interior, and assuming that $f(t,\sigma)$ is the given input datum for $(t,\sigma) \in (0,T) \times \Gamma_{\text{in}}$ and that the corresponding output datum is the thermal flux $A(\sigma)\nabla u(T_0,\sigma)\cdot \mathbf{n}(\sigma)$ measured at a given time T_0 for $\sigma \in \Gamma_{\text{out}}$, we prove that knowledge of all possible pairs of input-output data

$$(f, A\nabla u(T_0)\cdot \mathbf{n}_{|\Gamma_{\text{out}}})$$

determines uniquely the boundary spectral data of the underlying elliptic operator. Under suitable hypothesis on ρ, A, q, their identifiability is then proved. The same results hold when a mean value of the thermal flux is measured over a small interval of time.

Key words. uniqueness, inverse problem

AMS subject classifications. 35, 35R30

PII. S003614109936525X

1. Introduction and main results. We denote by $u(t,x)$ the temperature of a sufficiently smooth body Ω in $\mathbb{R}^N, N \geq 2$, at the time t and at the point $x \in \Omega, u_0$ is the initial temperature, f the temperature on $(0,T) \times \partial\Omega, \rho(x)$ the density, $q(x)$ the potential, and $A(x)$ the anisotropic thermal diffusion coefficient. We make the following regularity assumptions on the functions ρ, A, q:

$$\rho \in L^\infty(\Omega) \text{ and } \rho(x) \geq \beta \text{ for some constant } \beta > 0; \tag{1.1}$$

$A(x) = (a_{i\ell}(x))_{1\leq i,\ell\leq N}$ is a symmetric $N \times N$ matrix valued function in Ω satisfying the following conditions:

(i) $(a_{i\ell})_{1\leq i,\ell\leq N} \in C^{0,1}(\overline{\Omega})$, that is, there exists a constant $C > 0$ such that for every $x, y \in \overline{\Omega}$

$$|a_{i\ell}(x) - a_{i\ell}(y)| \leq C\,|x-y|\,, \quad i,\ell = 1,\ldots,N; \tag{1.2}$$

(ii) there exists a constant $\alpha > 0$ such that for every $x \in \overline{\Omega}$ and $\xi \in \mathbb{R}^N$

$$A(x)\xi\cdot\xi \geq \alpha\,|\xi|^2\,, \tag{1.3}$$

where $x \cdot y$ denotes the euclidean scalar product of two elements $x, y \in \mathbb{R}^N$;

$$q \in L^p(\Omega) \text{ for some } p > \frac{N}{2}. \tag{1.4}$$

Supposing ρ, A, q, u_0, and f assigned, then u solves the following heat equation, which we call the *direct problem*:

$$\left\{\begin{array}{rcll} \rho(x)\partial_t u - \operatorname{div}(A(x)\nabla u) + q(x)u &=& 0 & \text{in } (0,T)\times\Omega, \\ u(0) &=& u_0 & \text{in } \Omega, \\ u(t,\sigma) &=& f(t,\sigma) & \text{on } (0,T)\times\partial\Omega. \end{array}\right. \tag{1.5}$$

*Received by the editors December 20, 1999; accepted for publication (in revised form) August 1, 2000; published electronically January 5, 2001.
http://www.siam.org/journals/sima/32-5/36525.html

†Laboratoire de Mathématiques Appliquées (UMR7641), Université de Versailles, 45, Avenue des Etats-Unis, 78035 Versailles cedex, France (canuto@math.uvsq.fr, kavian@math.uvsq.fr).

It is well known that, under reasonable assumptions on the data, (1.5) has a unique solution, and that the thermal flux

$$A(\sigma)\nabla u(t,\sigma)\cdot\mathbf{n}(\sigma)$$

is well defined for $(t,\sigma)\in(0,T)\times\partial\Omega$.

In the present paper we are interested in the study of identifiability of the coefficients ρ, A, q in (1.5), when, assigning the temperature f on $(0,T)\times\partial\Omega$, one measures the corresponding thermal flux $A\nabla u(T_0)\cdot\mathbf{n}_{|\Gamma_{\rm out}}$ at a given time T_0 on a piece $\Gamma_{\rm out}$ of the boundary of Ω. More generally we may impose the input f on $(0,T)\times\Gamma_{\rm in}$, where $\Gamma_{\rm in}$ is a piece of the boundary of Ω, and we ask the following question: if we denote by Λ the so-called input-output map, that is,

$$\Lambda : f \longmapsto A\nabla u(T_0)\cdot\mathbf{n}_{|\Gamma_{\rm out}}, \tag{1.6}$$

where $T_0\in(0,T]$ is a given fixed time, and by Φ the nonlinear operator

$$\Phi : (\rho, A, q) \longmapsto \Lambda, \tag{1.7}$$

is Φ *injective*? We point out that to prove the injectivity of the operator Φ in (1.7), it is equivalent to show the uniqueness of the coefficients ρ, A, q in Ω from knowledge of all possible pairs of input-output data

$$(f, A\nabla u(T_0)\cdot\mathbf{n}_{|\Gamma_{\rm out}})$$

of the solution u of (1.5), that is, from an infinite number of measurements of the thermal flux $A\nabla u(T_0)\cdot\mathbf{n}_{|\Gamma_{\rm out}}$ at a given time T_0 on $\Gamma_{\rm out}$. We observe moreover that the thermal fluxes are measured at a given time T_0 *only*, instead of measuring it over a whole interval of time such as $[0,T_0]$. We study also the problem in which a *mean* value (over a small interval of time) of the thermal flux is measured.

Similar problems have been studied by many authors. Kravaris and Seinfeld [18] consider the problem of identifiability of the coefficient $a(x)\in C^1[0,\ell]$ in the one dimensional heat equation

$$\left\{\begin{array}{rcll} \partial_t u-\partial_x(a(x)\partial_x u) &=& 0 & \text{in } (0,\ell)\times(0,T),\\ u(0) &=& u_0 & \text{in } (0,\ell),\\ a(0)\partial_x u(t,0) &=& f(t) & \text{in } (0,T),\\ a(\ell)\partial_x u(t,\ell) &=& 0 & \text{in } (0,T)\end{array}\right. \tag{1.8}$$

from the additional measurement of the temperature $u(t,x_*)$ on $[0,T]$ at one end-point x_* of the interval $[0,\ell]$ when a single input f is assigned in (1.8). More precisely, they prove that if the temperature $u(t,x_*)$ is measured at the end-point $x_*=0$, where the rod is supposedly heated, then the coefficient $a(x)$ is uniquely determined in $[0,\ell]$ provided that the input f satisfies the following condition: for some $\varepsilon>0$, $f(t)$ is not identically zero in $(0,\varepsilon)$. On the other hand, if the temperature is measured at the other end-point $x_*=\ell$, where the rod is supposedly insulated, then in general $a(x)$ is not uniquely determined, except for symmetric coefficients, i.e., $a(\ell-x)=a(x)$, in which case uniqueness holds. These uniqueness results are obtained as an implication of identifiability of the coefficient $a(x)$ in the Sturm–Liouville problem

$$\left\{\begin{array}{rcll} -(a\varphi_k')' &=& \lambda_k\varphi_k & \text{in } (0,\ell),\\ \varphi_k'(0)=\varphi_k'(\ell) &=& 0,\\ \displaystyle\int_0^\ell |\varphi_k|^2\,dx &=& 1\end{array}\right.$$

when the sequence of eigenvalues $(\lambda_k)_{k=1}^{\infty}$ and eigenfunctions $(|\varphi_k(0)|)_{k=1}^{\infty}$ is known.

For $N \geq 2$, Isakov [14] proves the unique determination of the scalar coefficients $a(x)$ and $q(x), a \in W^{2,\infty}(\Omega)$ and $q \in L^{\infty}(\Omega)$, in the equation

$$\left\{\begin{array}{rcll} \partial_t u - \operatorname{div}(a(x)\nabla u) + q(x)u &=& 0 & \text{in } (0,T)\times\Omega, \\ u(0) &=& 0 & \text{in } \Omega, \\ u(t,\sigma) &=& f(t,\sigma) & \text{on } (0,T)\times\partial\Omega \end{array}\right. \tag{1.9}$$

from the measurement of the flux $\frac{\partial}{\partial \mathbf{n}} u$ on $[0,T]\times\partial\Omega$ when all possible input data f are assigned in (1.9). We note that in this case the flux is measured on a whole interval of time and on the whole boundary of Ω.

The first step in order to prove the injectivity of the operator Φ defined in (1.7) is to study the identifiability of the boundary spectral data for the underlying elliptic operator from the input-output map Λ. More precisely, let us denote by $(\lambda_k)_{k=1}^{\infty}$ and $(\varphi_k)_{k=1}^{\infty}$, respectively, the nondecreasing sequence of eigenvalues and the corresponding eigenfunctions of the Dirichlet problem

$$\left\{\begin{array}{rcll} -\operatorname{div}(A\nabla\varphi_k) + q\varphi_k &=& \lambda_k\rho\varphi_k & \text{in } \Omega, \\ \varphi_k &=& 0 & \text{on } \partial\Omega, \\ \displaystyle\int_{\Omega} |\varphi_k|^2 \rho dx &=& 1 & \end{array}\right.$$

and by $\mathrm{BSD}(\rho, A, q)$ the *boundary spectral data*, i.e.,

$$\mathrm{BSD}(\rho, A, q) := (\lambda_k, A\nabla\varphi_k \cdot \mathbf{n}_{|\Gamma_{\mathrm{in}}\cup\Gamma_{\mathrm{out}}})_{k=1}^{\infty}.$$

The question we ask is the following: does the input-output map Λ defined in (1.6) determine the boundary spectral data $\mathrm{BSD}(\rho, A, q)$ uniquely? The first result of the present paper is the following.

THEOREM 1.1. *Let $N \geq 2$ be an integer, Ω be a bounded domain in $\mathbb{R}^N$ of class $C^{1,1}$, and let $\Gamma_{\mathrm{in}}, \Gamma_{\mathrm{out}}$ be two relatively open pieces of $\partial\Omega$ such that $\Gamma_{\mathrm{in}} \cap \Gamma_{\mathrm{out}}$ has a nonempty interior. For $j \in \{0,1\}$, consider two sets of functions (ρ_j, A_j, q_j) satisfying conditions* (1.1)–(1.4). *For some fixed $u_{0j} \in L^2(\Omega)$, and $\varphi_j \in C([0,T]; H^{\frac{3}{2}}(\partial\Omega\backslash\Gamma_{\mathrm{in}}))$, let $u_j \in C^1([O,T], L^2(\Omega)) \cap C((O,T], H^2(\Omega))$ solve*

$$\left\{\begin{array}{rcll} \rho_j(x)\partial_t u_j - \operatorname{div}(A_j(x)\nabla u_j) + q_j(x)u_j &=& 0 & \textit{in } (0,T)\times\Omega, \\ u_j(0) &=& u_{0j} & \textit{in } \Omega, \\ u_j &=& \varphi_j & \textit{on } (0,T)\times\partial\Omega\backslash\Gamma_{\mathrm{in}}, \\ u_j &=& f & \textit{on } (0,T)\times\Gamma_{\mathrm{in}}. \end{array}\right. \tag{1.10}$$

We denote by

$$\Lambda_j(f) := A_j\nabla u_j(T_0)\cdot\mathbf{n}_{|\Gamma_{\mathrm{out}}} \tag{1.11}$$

the thermal fluxes measured at a given time $T_0 \in (0,T]$ on Γ_{out}. Suppose that one has

$$\Lambda_0(f) = \Lambda_1(f) \quad \textit{in } H^{\frac{1}{2}}(\Gamma_{\mathrm{out}}) \tag{1.12}$$

for all $f \in C([0,T]; H^{\frac{3}{2}}(\Gamma_{\mathrm{in}}))$ such that the $\operatorname{supp}(f(t,\cdot)) \subset \Gamma_{\mathrm{in}}$ for $t \in [0,T]$. Then the boundary spectral data $\mathrm{BSD}(\rho_j, A_j, q_j), j \in \{0,1\}$, coincide, that is, up to an appropriate choice of the eigenfunctions φ_{0k}, for all $k \geq 1$, one has

$$\lambda_{0k} = \lambda_{1k}, \quad \textit{and} \quad A_0\nabla\varphi_{0k}\cdot\mathbf{n}_{|\Gamma_{\mathrm{in}}\cup\Gamma_{\mathrm{out}}} = A_1\nabla\varphi_{1k}\cdot\mathbf{n}_{|\Gamma_{\mathrm{in}}\cup\Gamma_{\mathrm{out}}} \quad \textit{on } \Gamma_{\mathrm{in}}\cup\Gamma_{\mathrm{out}}.$$

(Here $\mathbf{n}(\sigma)$ denotes the outer unit normal at $\sigma \in \partial\Omega$.) We observe that the initial data u_{0j} and the boundary data φ_j in (1.10) are assigned arbitrarily. In fact this assumption corresponds to a realistic situation in which the data u_{0j}, φ_j are a priori unknown.

We point out that the conclusion of Theorem 1.1 remains valid if we replace hypothesis (1.12) by equality of the mean values of the thermal fluxes in the interval $[\tau_0 - T_0, T_0]$, that is, the following theorem holds.

THEOREM 1.2. *Let $0 < \tau_0 < T_0$ be given. Under the assumptions of Theorem* 1.1 *we suppose that*

$$\int_{T_0-\tau_0}^{T_0} A_0 \nabla u_0(t) \mathbf{n}_{|\Gamma_{\text{out}}} dt = \int_{T_0-\tau_0}^{T_0} A_1 \nabla u_1(t) \mathbf{n}_{|\Gamma_{\text{out}}} dt \quad \text{in } H^{\frac{1}{2}}(\Gamma_{\text{out}})$$

for all $f \in C([0,T]; H^{\frac{3}{2}}(\Gamma_{\text{in}}))$ such that the $\text{supp}(f(t,\cdot)) \subset \Gamma_{\text{in}}$ *for $t \in [0,T]$. Then the boundary spectral data* $\text{BSD}(\rho_j, A_j, q_j), j \in \{0,1\}$, *coincide, that is, up to an appropriate choice of the eigenfunctions φ_{0k}, for all $k \geq 1$, one has*

$$\lambda_{0k} = \lambda_{1k}, \quad \text{and} \quad A_0 \nabla \varphi_{0k} \cdot \mathbf{n}_{|\Gamma_{\text{in}} \cup \Gamma_{\text{out}}} = A_1 \nabla \varphi_{1k} \cdot \mathbf{n}_{|\Gamma_{\text{in}} \cup \Gamma_{\text{out}}} \quad \text{on } \Gamma_{\text{in}} \cup \Gamma_{\text{out}}.$$

Once the result of Theorem 1.1 is at hand, we can prove the injectivity of the operator Φ defined in (1.7) in the following three cases:

(i) given $q(x)$ and $A(x)$ of the form $A(x) = a(x) I_N$, where $a(x)$ is a scalar-valued function and I_N is the identity matrix, we identify $\rho(x)$ and $a(x)$ by supposing that $\Gamma_{\text{in}} = \Gamma_{\text{out}} = \partial\Omega$;

(ii) given $A(x) = a(x) I_N$, we identify $\rho(x)$ and $q(x)$ by supposing that $\Gamma_{\text{in}} = \Gamma_{\text{out}} = \partial\Omega$;

(iii) given $\rho(x)$ and $A(x) = a(x) I_N$, we identify $q(x)$ by supposing that $\Gamma_{\text{in}} \cup \Gamma_{\text{out}} = \partial\Omega$.

In what follows we suppose that the coefficients ρ, a, q satisfy the following regularity assumptions:

$$\rho \in L^\infty(\Omega) \text{ and } \rho \geq \beta > 0 \text{ for some constant } \beta; \tag{1.13}$$

$$a \geq \alpha > 0 \text{ for some constant } \alpha, \tag{1.14}$$

when $N = 2$

$$a \in W^{1,p}(\Omega) \text{ for some } p > 2, \tag{1.15}$$

when $N \geq 3$

$$a \in C^{1,\frac{1}{2}+\varepsilon}(\overline{\Omega}); \tag{1.16}$$

when $N = 2$

$$q \in L^p(\Omega) \text{ for some } p > 1 \text{ and } q \geq -\mu_1, \tag{1.17}$$

when $N \geq 3$

$$q \in L^p(\Omega) \text{ for some } p > \frac{N}{2}. \tag{1.18}$$

Here μ_1 denotes the first eigenvalue of $-\Delta$ with Dirichlet boundary condition.

In the following theorems we prove the identifiability of the coefficients ρ, a, q in cases (i)–(iii) mentioned above, under hypothesis (1.13)–(1.18). We begin by proving case (i), that is, given q, we prove the identifiability of ρ and a.

THEOREM 1.3. *Under the assumptions of Theorem* 1.1, *let* $\Gamma_{\text{in}} = \Gamma_{\text{out}} = \partial\Omega$, *and, for* $j \in \{0,1\}$, *let* (ρ_j, a_j, q) *be two sets of functions satisfying conditions* (1.13)–(1.18) *such that, when* $N = 2, q \equiv 0$ *in* Ω. *We denote by* u_j *the solutions of problems* (1.10) *when* $(\rho_j, A_j, q_j) := (\rho_j, a_j, q)$. *Suppose that*

$$\Lambda_0(f) = \Lambda_1(f) \quad \text{in } H^{\frac{1}{2}}(\partial\Omega) \tag{1.19}$$

for all $f \in C([0,T]; H^{\frac{3}{2}}(\partial\Omega))$, *where* $\Lambda_j(f)$ *are defined in* (1.11). *Then* $\rho_0 = \rho_1$ *in* Ω *and* $a_0 = a_1$ *in* $\overline{\Omega}$.

We point out that in this case the measurements are global, since Γ_{in} and Γ_{out} coincide with the whole boundary of Ω. In the following theorem, given a, we prove the identifiability of ρ and q.

THEOREM 1.4. *Under the assumptions of Theorem* 1.1, *let* $\Gamma_{\text{in}} = \Gamma_{\text{out}} = \partial\Omega$, *and, for* $j \in \{0,1\}$, *let* (ρ_j, a, q_j) *be two sets of functions satisfying conditions* (1.13)–(1.18) *such that, when* $N = 2, a \equiv 1$ *in* $\overline{\Omega}$. *We denote by* u_j *the solutions of problems* (1.10) *when* $(\rho_j, A_j, q_j) := (\rho_j, a, q_j)$. *Suppose that*

$$\Lambda_0(f) = \Lambda_1(f) \quad \text{in } H^{\frac{1}{2}}(\partial\Omega) \tag{1.20}$$

for all $f \in C([0,T]; H^{\frac{3}{2}}(\partial\Omega))$, *where* $\Lambda_j(f)$ *are defined in* (1.11). *Then* $\rho_0 = \rho_1$ *and* $q_0 = q_1$ *in* Ω.

In the following theorem we give a simplified proof of a result of Nachman, Sylvester, and Uhlmann [22] concerning an N-dimensional Borg–Levinson theorem, which we will use below to prove the identifiability of q in case (iii).

THEOREM 1.5. *Let* $N \geq 2$ *be an integer,* Ω *be a bounded domain in* $\mathbb{R}^N$ *of class* $C^{1,1}$, *and, for* $j \in \{0,1\}$, *let* (ρ, a, q_j) *be a set of functions satisfying conditions* (1.13)–(1.18) *such that, when* $N = 2, a \equiv 1$ *in* $\overline{\Omega}$. *We denote by* $(\lambda_{jk})_{k=1}^{\infty}$ *and* $(\varphi_{jk})_{k=1}^{\infty}$, *respectively, the eigenvalues and the corresponding eigenfunctions of the following problems:*

$$\begin{cases} -\operatorname{div}(a\nabla\varphi_{jk}) + q_j\varphi_{jk} = \lambda_{jk}\rho\varphi_{jk} & \text{in } \Omega, \\ \varphi_{jk} = 0 & \text{on } \partial\Omega, \\ \displaystyle\int_\Omega |\varphi_{jk}|^2 \rho dx = 1. \end{cases} \tag{1.21}$$

Suppose that, for all $k \geq 1$,

$$\lambda_{0k} = \lambda_{1k} \quad \text{and} \quad \psi_{0k} = \psi_{1k} \quad \text{on } \partial\Omega,$$

where $\psi_{jk} := a\frac{\partial}{\partial \mathbf{n}}\varphi_{jk|\partial\Omega}$. *Then* $q_0 = q_1$ *in* Ω.

(Note that the above result is slightly more general than the one given in [21], although the idea of its proof is essentially the same.) As a consequence of Theorems 1.1 and 1.5, given ρ and a, we prove the identifiability of q in case (iii).

THEOREM 1.6. *Under the assumptions of Theorem* 1.1, *let* $\Gamma_{\text{in}} \cup \Gamma_{\text{out}} = \partial\Omega$, *and, for* $j \in \{0,1\}$, *let* (ρ, a, q_j) *be a set of functions satisfying conditions* (1.13)–(1.18) *such that, when* $N = 2, a \equiv 1$ *in* $\overline{\Omega}$. *We denote by* u_j *the solutions of problems* (1.10) *when* $(\rho_j, A_j, q_j) := (\rho, a, q_j)$. *Suppose that*

$$\Lambda_0(f) = \Lambda_1(f) \quad \text{in } H^{\frac{1}{2}}(\Gamma_{\text{out}}) \tag{1.22}$$

for all $f \in C([0,T]\,; H^{\frac{3}{2}}(\Gamma_{\rm in}))$ *such that the* $\mathrm{supp}(f(t,\cdot)) \subset \Gamma_{\rm in}$ *for* $t \in [0,T]$*, where* $\Lambda_j(f)$ *are defined in* (1.11)*. Then* $q_0 = q_1$ *in* Ω.

We point out that in this case the measurements are local, since $\Gamma_{\rm in}$ and $\Gamma_{\rm out}$ are two relatively open pieces of the boundary of Ω.

REMARK 1.7. We emphasize that, as in Theorem 1.1, the conclusions of Theorems 1.3, 1.4, and 1.6 remain valid if we replace hypotheses (1.19), (1.20), (1.22), respectively, by equality of the mean values of the thermal fluxes in the interval $[\tau_0 - T_0, T_0]$. □

The remainder of the paper is organized as follows: in section 2 we gather some preliminary results and the notations used throughout; in section 3 we prove the main results, that is, Theorem 1.1 and Theorem 1.2; in section 4 we prove Theorem 1.5; in section 5 we prove Theorems 1.3, 1.4, and 1.6; finally in section 6 we consider the same problem in the one dimensional setting.

2. Preliminary results and notations. We give now a list of notations which are used throughout the paper.

We denote by Ω a bounded domain in $\mathbb{R}^N, N \geq 2$, with boundary $\partial\Omega$ of class $C^{1,1}$. The functions ρ, A, q satisfy assumptions (1.1)–(1.4).

The thermal flux of the solution u to (1.5) at a point $\sigma \in \partial\Omega$ is denoted by

$$\psi(t,\sigma) := A(\sigma)\nabla u(t,\sigma)\cdot \mathbf{n}(\sigma),$$

where $x \cdot y$ denotes the euclidean scalar product in $\mathbb{R}^N$ and $\mathbf{n}(\sigma)$ is the outer unit normal at $\sigma \in \partial\Omega$.

We denote by $(L, \mathcal{D}(L))$ the elliptic operator

$$Lu := -\mathrm{div}(A\,(x)\,\nabla u) + q(x)u, \tag{2.1}$$

with domain $\mathcal{D}(L) := \left\{u \in H_0^1(\Omega);\ Lu \in L^2_\rho(\Omega)\right\}$, where

$$L^2_\rho(\Omega) := \left\{ f \in L^2(\Omega);\ \int_\Omega |f|^2\,\rho dx < \infty \right\},$$

equipped with the scalar product

$$(f \mid g) := \int_\Omega fg\rho dx.$$

Note that $L^2(\Omega) = L^2_\rho(\Omega)$ with equivalent norms. Actually, when $A \in C^{0,1}(\overline{\Omega})$ and $\partial\Omega$ is of class $C^{1,1}$, then

$$\mathcal{D}(L) = H^2(\Omega) \cap H_0^1(\Omega).$$

The operator L possesses a sequence of eigenvalues $(\lambda_k)_{k=1}^\infty$ (which we suppose in a nondecreasing order) and of corresponding eigenfunctions $(\varphi_k)_{k=1}^\infty$ satisfying

$$\left\{ \begin{array}{rcll} -\mathrm{div}(A\nabla\varphi_k) + q\varphi_k & = & \lambda_k\rho\varphi_k & \text{in } \Omega, \\ \varphi_k & = & 0 & \text{on } \partial\Omega, \\ \int_\Omega |\varphi_k|^2\,\rho dx & = & 1 & \end{array} \right. \tag{2.2}$$

which form a Hilbert basis of $L^2_\rho(\Omega)$. It is also known that the domain $\mathcal{D}(L)$ can be characterized by

$$\mathcal{D}(L) = \left\{ u \in L^2(\Omega);\ \sum_{k=1}^\infty \lambda_k^2\,|(u \mid \varphi_k)|^2 < +\infty \right\}, \tag{2.3}$$

where $(\cdot \mid \cdot)$ is the scalar product in $L^2_\rho(\Omega)$. The flux of the eigenfunction φ_k on $\partial\Omega$ will be denoted by ψ_k, that is,

$$\psi_k := A\nabla\varphi_k \cdot \mathbf{n}_{|\partial\Omega}. \tag{2.4}$$

We denote by m_k the geometric multiplicity of λ_k. We recall also the asymptotic behavior of the eigenvalues λ_k:

$$\lambda_k \sim C_0 k^{\frac{2}{N}} \quad \text{as } k \to +\infty,$$

where the constant C_0 depends on ρ, A, Ω, N (see Courant and Hilbert [10, pp. 442–443]). Moreover there exist two positive constants C_1, C_2 such that, for all $k \geq 1$, one has

$$C_1\lambda_k^2 \leq \|\varphi_k\|^2_{H^2(\Omega)} \leq C_2\lambda_k^2. \tag{2.5}$$

We shall need the following result concerning the linear independence, or linear dependence, of the family $(\psi_k)_{k\geq 1}$ defined above in (2.4). In general these functions are not linearly independent. However, one can show that the normal derivatives of the eigenfunctions corresponding to a given eigenvalue λ_{k_0} are actually independent. More precisely, if λ_k is an eigenvalue of L having multiplicity $m_k \geq 1$, let us denote by $\varphi_{k,i}$ for $1 \leq i \leq m_k$ the eigenfunctions corresponding to the eigenvalue λ_k which form a Hilbert basis of the kernel $N(L - \lambda_k I)$. We may state the following lemma.

LEMMA 2.1. *Under the above conditions* (1.1)–(1.4) *on the coefficients* ρ, A, q, *let* $k \geq 1$ *be fixed. If* λ_k *is an eigenvalue of multiplicity* $m_k \geq 1$ *and* Γ *is a relatively open piece of* $\partial\Omega$, *then the dimension of the subspace spanned in* $L^2(\Gamma)$ *by* $(\psi_{k,i})_{1\leq i\leq m_k}$ *is exactly* m_k.

Proof of Lemma 2.1. Indeed if there exists $(c_i)_{i=1}^{m_k} \in \mathbb{R}^{m_k}$ such that

$$\sum_{i=1}^{m_k} c_i\psi_{k,i} = 0 \qquad \text{on } \Gamma,$$

then setting $\varphi := \sum_{i=1}^{m_k} c_i\varphi_{k,i}$, one checks that

$$L\varphi = \lambda_k\varphi \quad \text{in } \Omega, \qquad \varphi = 0 \quad \text{on } \partial\Omega, \quad A\nabla\varphi \cdot \mathbf{n}_{|\Gamma} = 0 \quad \text{on } \Gamma.$$

So the unique continuation principle at the boundary implies that $\varphi \equiv 0$ in Ω (see Adolfsson and Escauriaza [1]). Due to the fact that the functions $\varphi_{k,i}$ are linearly independent, we conclude that $c_i = 0$ for $1 \leq i \leq m_k$. The proof is complete. ∎

From this we conclude the following.

LEMMA 2.2. *Under the assumptions of Lemma* 2.1, *let* Γ_{in} *and* Γ_{out} *be two relatively open pieces of* $\partial\Omega$. *For a fixed* $k \geq 1$, *consider the function* Θ_k *defined by*

$$\Theta_k(\sigma', \sigma) := \sum_{i=1}^{m_k} \psi_{k,i}(\sigma')\psi_{k,i}(\sigma) \quad \text{on } \Gamma_{\text{in}} \times \Gamma_{\text{out}}.$$

Then $\Theta_k(\sigma', \sigma)$ *is not identically zero on any relatively open subset of* $\Gamma_{\text{in}} \times \Gamma_{\text{out}}$.

Proof of Lemma 2.2. By contradiction, let Γ_1 be a relatively open piece of Γ_{in}, Γ_2 a relatively open piece of Γ_{out}, and

$$\Theta_k(\sigma', \sigma) \equiv 0 \quad \text{on } \Gamma_1 \times \Gamma_2. \tag{2.6}$$

By the previous Lemma 2.1, (2.6) implies that $\psi_{k,i} \equiv 0$ on Γ_1 for $i = 1, \dots, m_k$, and so by the unique continuation result it follows that $\varphi_{k,i} \equiv 0$ on Ω, which leads to a contradiction. □

We shall also need the following algebraic lemma.

LEMMA 2.3. *Let $m \geq 1$ and $n \geq 1$ be two arbitrary integers, Z a nonempty set, X, Y two subsets of Z, $f_i : X \cup Y \to \mathbb{R}$ (for $1 \leq i \leq m$) and $g_\ell : X \cup Y \to \mathbb{R}$ (for $1 \leq \ell \leq n$) functions such that*

$$\sum_{i=1}^{m} f_i(x) f_i(y) = \sum_{\ell=1}^{n} g_\ell(x) g_\ell(y) \quad \text{for } x, y \in X \times Y. \tag{2.7}$$

Assume moreover that

(i) *$\{f_1, \dots, f_m\}$ (resp., $\{g_1, \dots, g_n\}$) are linearly independent in $X \cup Y$;*

(ii) *$X \cap Y$ contains infinitely many points;*

(iii) *f_i, for $1 \leq i \leq m$, (resp., g_ℓ, for $1 \leq \ell \leq n$) are not identically zero in $X \cap Y$.*

Then $m = n$, and denoting

$$F(x) := \begin{pmatrix} f_1(x) \\ \vdots \\ f_m(x) \end{pmatrix} \quad \text{and} \quad G(x) := \begin{pmatrix} g_1(x) \\ \vdots \\ g_n(x) \end{pmatrix},$$

there exists an $m \times m$ orthogonal matrix M such that for all $z \in X \cup Y$ one has $F(z) = MG(z)$.

(Recall that by an orthogonal matrix M we mean $MM^* = M^*M = I_m$.)

Proof of Lemma 2.3. Let us denote by V_0 (resp., V_1) the space spanned by $\{f_1, \dots, f_m\}$ (resp., $\{g_1, \dots, g_n\}$). As f_1 is not identically zero in $X \cap Y$, there exists $x_1 \in X \cap Y$ such that $f_1(x_1) \neq 0$. Then, f_2 being independent of f_1, and $X \cap Y$ containing infinitely many points, there exists $x_2 \in X \cap Y$ such that

$$\det \begin{pmatrix} f_1(x_1) & f_2(x_1) \\ f_1(x_2) & f_2(x_2) \end{pmatrix} \neq 0.$$

By induction one sees that we may find points $x_1, x_2, \dots, x_m$ in $X \cap Y$ such that the $m \times m$ matrix

$$P := \begin{pmatrix} f_1(x_1) & f_2(x_1) & \cdots & f_m(x_1) \\ f_1(x_2) & f_2(x_2) & \cdots & f_m(x_2) \\ \vdots & \vdots & \vdots & \vdots \\ f_1(x_m) & f_2(x_m) & \cdots & f_m(x_m) \end{pmatrix}$$

is invertible. So, setting $x = x_j$ in (2.7), it follows that $PF(y) = \widetilde{P}G(y)$ in Y, where $\widetilde{P}$ is the following $m \times n$ matrix:

$$\widetilde{P} := \begin{pmatrix} g_1(x_1) & g_2(x_1) & \cdots & g_n(x_1) \\ g_1(x_2) & g_2(x_2) & \cdots & g_n(x_2) \\ \vdots & \vdots & \vdots & \vdots \\ g_1(x_m) & g_2(x_m) & \cdots & g_n(x_m) \end{pmatrix}.$$

From this it follows that $F(y) = P^{-1}\widetilde{P}G(y)$ for all $y \in Y$, where P^{-1} is the inverse matrix of P.

Similarly, changing the role of the variables x and y, we obtain that $F(x) = P^{-1}\widetilde{P}G(x)$ in X, that is,

$$F(z) = MG(z) \quad \text{in } X \cup Y,$$

where $M := P^{-1}\widetilde{P}$. Therefore, recalling that the functions $\{f_1, \dots, f_m\}$ and $\{g_1, \dots, g_n\}$ are linearly independent in $X \cup Y$, it follows that $V_0 \subseteq V_1$, that is, $m \leq n$. In the same way one may prove that $n \leq m$, and so we conclude that $m = n$.

Finally we prove that the matrix $M = P^{-1}\widetilde{P}$ is orthogonal. Indeed, recalling that $F(z) = MG(z)$ for all $z \in X \cup Y$ and using (2.7), we obtain

$$(M^*M - I_m)G(x) \cdot G(y) = 0 \quad \text{in } X \times Y,$$

where $a \cdot b$ denotes the euclidean scalar product in $\mathbb{R}^m$, M^* is the transpose matrix of M, and I_m is the $m \times m$ identity matrix. Since the functions $\{g_1, \dots, g_m\}$ are linearly independent in $X \cup Y$ it follows that $M^*M = MM^* = I_m$, that is, M is orthogonal. The proof is complete. ∎

3. Proof of Theorems 1.1 and 1.2. We begin this section with two lemmas which will be needed later on. First we give a representation formula for the solution u of problem (1.5).

LEMMA 3.1. *Let Ω be a $C^{1,1}$ bounded domain in $\mathbb{R}^N$, (ρ, A, q) be a set of functions satisfying assumptions* (1.1)–(1.4)*, and $f \in C([0,T]\,; H^{\frac{3}{2}}(\partial\Omega))$. Then the problem*

$$\text{(3.1)} \qquad \begin{cases} \rho(x)\partial_t u - \operatorname{div}(A(x)\nabla u) + q(x)u &= 0 & \text{in } (0,T)\times\Omega, \\ u(0) &= 0 & \text{in } \Omega, \\ u(t,\sigma) &= f(t,\sigma) & \text{on } (0,T)\times\partial\Omega \end{cases}$$

has a unique solution $u \in C((0,T]; H^2(\Omega)) \cap C^1([0,T]\,; L^2(\Omega))$. Moreover, u can be written in the following Fourier expansion:

$$\text{(3.2)} \qquad u(t) = \sum_{k=1}^{\infty} \alpha_k(t)\varphi_k \quad \text{in } L^2_\rho(\Omega),$$

where $\alpha_k(t)$ are the Fourier coefficients of $u(t)$.

Proof of Lemma 3.1. By the classical theory of semigroups (see, for example, Cazenave and Haraux [7]) we know that there exists a unique solution of (3.1) such that $u \in C((0,T]; H^2(\Omega)) \cap C^1([0,T]\,; L^2(\Omega))$. Now, since $(\varphi_k)_{k=1}^{\infty}$ defined in (2.2) is a Hilbert basis in $L^2_\rho(\Omega)$, we can write the solution u in the following Fourier expansion:

$$u(t) = \sum_{k=1}^{\infty} \alpha_k(t)\varphi_k \quad \text{in } L^2_\rho(\Omega),$$

where the coefficient $\alpha_k(t) := (u(t) \mid \varphi_k)$ solves

$$\text{(3.3)} \qquad \begin{cases} \frac{d}{dt}\alpha_k(t) + \lambda_k\alpha_k(t) &= -\int_{\partial\Omega} \psi_k(\sigma')f(t,\sigma')d\sigma' \quad \text{in } (0,T), \\ \alpha_k(0) &= 0. \end{cases}$$

Here $\frac{d}{dt}\alpha_k(t)$ denotes the derivative of $\alpha_k(t)$, and $\psi_k(\sigma') := A(\sigma')\nabla\varphi_k(\sigma') \cdot \mathbf{n}(\sigma')$. A direct calculation gives

$$\text{(3.4)} \qquad \alpha_k(t) = -\int_0^t \int_{\partial\Omega} \psi_k(\sigma')e^{-\lambda_k(t-\tau)}f(\tau,\sigma')d\sigma'd\tau.$$

Lemma 3.1 is proved. □

In the following lemma we give, under a suitable choice of the Dirichlet boundary datum f, a representation formula for the thermal flux $A\nabla u(T_0)\cdot\mathbf{n}_{|\partial\Omega}$ at a given time T_0 on $\partial\Omega$.

LEMMA 3.2. *Under the assumptions of Lemma* 3.1*, we suppose that the Dirichlet boundary datum f in* (3.1) *satisfies the following condition:*

$$f \equiv 0 \quad \text{on } [T_0-\varepsilon_0, T_0]\times\partial\Omega,$$

where $T_0 \in (0,T]$ and $0<\varepsilon_0<T_0$. Then the thermal flux $A\nabla u(T_0)\cdot\mathbf{n}_{|\partial\Omega}$ at a given time T_0 on $\partial\Omega$ can be written in the following form:

$$A(\sigma)\nabla u(T_0,\sigma)\cdot\mathbf{n}(\sigma) = -\int_0^{T_0-\varepsilon_0}\int_{\partial\Omega}\sum_{k=1}^{\infty}\Psi_k(\sigma',\sigma,T_0-\tau)f(\tau,\sigma')d\sigma' d\tau$$

for almost everywhere (a.e.) $\sigma\in\partial\Omega$, where $\Psi_k(\sigma',\sigma,\tau) := \psi_k(\sigma')\,\psi_k(\sigma)e^{-\lambda_k\tau}$.

Proof of Lemma 3.2. We divide the proof into three steps.

Step 1. In this step we prove that, under a suitable choice of the Dirichlet datum f in (3.1), the series (3.2) converges at the time T_0, say, for instance, in $H^2(\Omega)$, so that we can take the trace of $\nabla u(T_0)_{|\partial\Omega}$ on $\partial\Omega$. In fact, by Lemma 3.1 we know that $u\in C((0,T];H^2(\Omega))\cap C^1([0,T];L^2(\Omega))$. In particular, choosing the Dirichlet datum f in such a way that $f\equiv 0$ on $[T_0-\varepsilon_0,T_0]\times\partial\Omega$, it follows that $u(T_0)\in\mathcal{D}(L)$, where $\mathcal{D}(L) := H^2(\Omega)\cap H_0^1(\Omega)$. Then, using the characterization (2.3) of the domain $\mathcal{D}(L)$, it follows that

$$\sum_{k=1}^{\infty}\lambda_k^2\,|\alpha_k(T_0)|^2 < \infty,$$

where $\alpha_k(T_0) = (u(T_0)\mid\varphi_k)$. Now, setting $u_m(T_0) := \sum_{k=1}^m \alpha_k(T_0)\varphi_k$, then $u_m(T_0)\to u(T_0)$ in $L^2_\rho(\Omega)$ as $m\to+\infty$; moreover, it is easy to verify that

$$Lu_m(T_0) - Lu(T_0) \to 0 \quad \text{in } L^2_\rho(\Omega).$$

So, by the estimate $\|u_m(T_0)\|_{H^2(\Omega)} \le C\,\|Lu_m(T_0)\|_{L^2_\rho(\Omega)}$, it follows that $(u_m(T_0))_{m=1}^\infty$ is a Cauchy sequences in $H^2(\Omega)$, and therefore

$$u(T_0) = \sum_{k=1}^{\infty}\alpha_k(T_0)\,\varphi_k \quad \text{in } H^2(\Omega). \tag{3.5}$$

Step 2. In this step we prove that the conormal derivative of the series (3.5) coincides with the series of the conormal derivative in $H^{\frac12}(\partial\Omega)$. In fact, it is well known that the trace operator $\gamma : u\to\frac{\partial}{\partial\mathbf{n}}u_{|\partial\Omega}$ is continuous from $H^2(\Omega)$ to $H^{\frac12}(\partial\Omega)$ (see, for example, Lions and Magenes [20]). Then one has, in the sense of $H^{\frac12}(\partial\Omega)$,

$$\frac{\partial}{\partial\mathbf{n}}u(T_0)_{|\partial\Omega} = -\sum_{k=1}^{\infty}\left(\int_0^{T_0-\varepsilon_0}\int_{\partial\Omega}\psi_k(\sigma')e^{-\lambda_k(T_0-\tau)}f(\tau,\sigma')d\sigma' d\tau\right)\frac{\partial}{\partial\mathbf{n}}\varphi_{k|\partial\Omega}$$

and

$$A(\sigma)\nabla u(T_0,\sigma)\cdot\mathbf{n}(\sigma) = -\sum_{k=1}^{\infty}\int_0^{T_0-\varepsilon_0}\int_{\partial\Omega}\Psi_k(\sigma',\sigma,T_0-\tau)f(\tau,\sigma')d\sigma' d\tau \tag{3.6}$$

for a.e. $\sigma \in \partial\Omega$, where $\Psi_k(\sigma', \sigma, \tau) := \psi_k(\sigma')\,\psi_k(\sigma)e^{-\lambda_k \tau}$.

Step 3. In this step we prove that we can commute the series sign with the integral signs in the right-hand side of (3.6).

By Fubini's theorem it is sufficient, for example, to show that

$$\mathcal{I} := \int_0^{T_0-\varepsilon_0} \sum_{k=1}^{\infty} \int_{\partial\Omega} \left|\psi_k(\sigma') e^{-\lambda_k(T_0-\tau)} f(\tau,\sigma')\right| \, \|\psi_k\|_{H^{\frac{1}{2}}(\partial\Omega)} d\sigma' d\tau < \infty.$$

In fact, denoting by $\langle\cdot,\cdot\rangle$ the duality $H^{-\frac{3}{2}}(\partial\Omega), H^{\frac{3}{2}}(\partial\Omega)$, we derive that

$$\begin{aligned}
\mathcal{I} &\le \int_0^{T_0-\varepsilon_0} \sum_{k=1}^{\infty} |\langle \psi_k, f(\tau)\rangle| \, \|\psi_k\|_{H^{\frac{1}{2}}(\partial\Omega)} \, e^{-\lambda_k \varepsilon_0} d\tau \\
&\le \int_0^{T_0-\varepsilon_0} \|f(\tau)\|_{H^{\frac{3}{2}}(\partial\Omega)} \, d\tau \sum_{k=1}^{\infty} \|\psi_k\|_{H^{-\frac{3}{2}}(\partial\Omega)} \, \|\psi_k\|_{H^{\frac{1}{2}}(\partial\Omega)} \, e^{-\lambda_k \varepsilon_0} \\
&\le C \sum_{k=1}^{\infty} \lambda_k^2 e^{-\lambda_k \varepsilon_0},
\end{aligned}$$

where the last inequality is obtained upon using the fact that, by the trace theorem and (2.5), we have

$$\begin{aligned}
\|\psi_k\|_{H^{-\frac{3}{2}}(\partial\Omega)} &\le \|\psi_k\|_{H^{\frac{1}{2}}(\partial\Omega)} \\
&\le C \, \|\varphi_k\|_{H^2(\Omega)} \le C\lambda_k.
\end{aligned}$$

Note that $\sum_{k=1}^{\infty} \lambda_k^2 e^{-\lambda_k \varepsilon_0} < \infty$ since we know that $\lambda_k \sim k^{\frac{2}{N}}$, as $k \to +\infty$. Therefore we may write (3.6) as

$$A(\sigma)\nabla u(T_0,\sigma)\cdot \mathbf{n}(\sigma) = -\int_0^{T_0-\varepsilon_0} \int_{\partial\Omega} \sum_{k=1}^{\infty} \Psi_k(\sigma', \sigma, T_0-\tau) f(\tau,\sigma') d\sigma' d\tau. \tag{3.7}$$

The proof of Lemma 3.2 is complete. □

LEMMA 3.3. *Under the assumptions of Theorem* 1.1, *let* $\overline{u_j}, j \in \{0,1\}$, *be solutions of* (1.10), *with initial data* $\overline{u_{0j}} \equiv 0$ *in* Ω, *and boundary data* $\overline{\varphi_j} \equiv 0$ *on* $(0,T)\times\partial\Omega\backslash\Gamma_{\text{in}}$. *If we denote by* $\overline{\Lambda_j}(f) := A_j\nabla\overline{u_j}(T_0)\cdot\mathbf{n}_{|\Gamma_{\text{out}}}$ *the output data of the solutions* $\overline{u_j}$, *then*

$$\overline{\Lambda_0}(f) = \overline{\Lambda_1}(f)$$

for all $f \in C([0,T]; H^{\frac{3}{2}}(\Gamma_{\text{in}}))$ *such that the* $\text{supp}(f(t,\cdot)) \subset \Gamma_{\text{in}}$ *for* $t \in [0,T]$.

Proof of Lemma 3.3. Indeed, following Rakesh and Symes [25], setting $\overline{u_j}(t,x) := u_j(t,x) - v_j(t,x)$, where v_j are solutions of (1.10) with data $v_j(0) = u_{0j}$ in $\Omega, v_j = \varphi_j$ on $(0,T)\times\partial\Omega\backslash\Gamma_{\text{in}}$, and $v_j \equiv 0$ on $(0,T)\times\Gamma_{\text{in}}$, one checks that $\overline{u_j}$ are solutions of (1.10) with initial data $\overline{u_{0j}} \equiv 0$ in Ω and boundary data $\overline{\varphi_j} \equiv 0$ on $(0,T)\times\partial\Omega\backslash\Gamma_{\text{in}}$. So, if we denote by $\overline{\Lambda_j}(f) := A_j\nabla\overline{u_j}(T_0)\cdot\mathbf{n}_{|\Gamma_{\text{out}}}$ the output data of the solutions $\overline{u_j}$, it follows that

$$\overline{\Lambda_j}(f) = \Lambda_j(f) - \Lambda_j(0),$$

and therefore $\overline{\Lambda_0}(f) = \overline{\Lambda_1}(f)$ for all $f \in C([0,T]\,; H^{\frac{3}{2}}(\Gamma_{\text{in}}))$ such that the support $\text{supp}(f(t,\cdot)) \subset \Gamma_{\text{in}}$ for $t \in [0,T]$.

The proof is complete. □

In what follows we will suppose that in (1.10) the initial data $u_{0j} \equiv 0$ in Ω and the boundary data $\varphi_j \equiv 0$ on $(0,T) \times \partial\Omega \backslash \Gamma_{\rm in}$.

LEMMA 3.4. *Under the assumptions of Theorem* 1.1, *for all* $k \geq 1$, *we have*

$$\lambda_{0k} = \lambda_{1k},$$

and

$$\sum_{i=1}^{m_{0k}} \psi_{0k,i}(\sigma')\,\psi_{0k,i}(\sigma) = \sum_{\ell=1}^{m_{1k}} \psi_{1k,\ell}(\sigma')\,\psi_{1k,\ell}(\sigma) \quad \text{on } \Gamma_{\rm in} \times \Gamma_{\rm out}.$$

(As we have mentioned in section 2, m_{jk} is the multiplicity of the eigenvalue λ_{jk}.)

Proof of Lemma 3.4. We recall that by hypothesis we have

$$A_0 \nabla u_0(T_0) \cdot \mathbf{n}_{|\Gamma_{\rm out}} = A_1 \nabla u_1(T_0) \cdot \mathbf{n}_{|\Gamma_{\rm out}} \quad \text{in } H^{\frac{1}{2}}(\Gamma_{\rm out}) \tag{3.8}$$

for all $f \in C([0,T]; H^{\frac{3}{2}}(\Gamma_{\rm in}))$ such that the support $\mathrm{supp}(f(t,\cdot)) \subset \Gamma_{\rm in}$ for $t \in [0,T]$. By Lemma 3.2, when $u := u_j, j \in \{0,1\}$, we know that if the input f is chosen in such a way that $f \equiv 0$ on $[T_0 - \varepsilon_0, T_0] \times \Gamma_{\rm in}$, then the fluxes $A_j \nabla u_j(T_0) \cdot \mathbf{n}_{|\Gamma_{\rm out}}$ on $\Gamma_{\rm out}$ can be written in the following form:

$$A_j(\sigma) u(T_0,\sigma) \cdot \mathbf{n}(\sigma) = -\int_0^{T_0-\varepsilon_0} \int_{\Gamma_{\rm in}} \sum_{k=1}^{\infty} \Psi_{jk}(\sigma',\sigma,T_0-\tau) f(\tau,\sigma') d\sigma' d\tau$$

for a.e. $\sigma \in \Gamma_{\rm out}$, where

$$\Psi_{jk}(\sigma',\sigma,\tau) := \psi_{jk}(\sigma')\,\psi_{jk}(\sigma)\,e^{-\lambda_{jk}\tau}. \tag{3.9}$$

Then, from (3.8), it follows that

$$\int_0^{T_0-\varepsilon_0} \int_{\Gamma_{\rm in}} \sum_{k=1}^{\infty} (\Psi_{0k}(\sigma',\sigma,T_0-\tau) - \Psi_{1k}(\sigma',\sigma,T_0-\tau)) f(\tau,\sigma') d\sigma' d\tau = 0 \tag{3.10}$$

for a.e. $\sigma \in \Gamma_{\rm out}$. In particular we may assume that the input $f(\tau,\sigma') \equiv 0$ for $\tau \neq [T' - \varepsilon', T' + \varepsilon']$ and $\sigma' \in \Gamma_{\rm in}$, where T' is a fixed time, $T' \in (0, T_0 - \varepsilon_0)$, and $0 < \varepsilon' < T'$. Then (3.10) becomes

$$\int_{T'-\varepsilon'}^{T'+\varepsilon'} \int_{\Gamma_{\rm in}} \sum_{k=1}^{\infty} (\Psi_{0k}(\sigma',\sigma,T_0-\tau) - \Psi_{1k}(\sigma',\sigma,T_0-\tau)) f(\tau,\sigma') d\sigma' d\tau = 0$$

for all such functions f. Hence it follows that

$$\sum_{k=1}^{\infty} \Psi_{0k}(\sigma',\sigma,\tau) = \sum_{k=1}^{\infty} \Psi_{1k}(\sigma',\sigma,\tau) \quad \text{on } \Gamma_{\rm in} \times \Gamma_{\rm out}$$

for all $\tau \in [T' - \varepsilon', T' + \varepsilon']$. Therefore, by the unique continuation principle for analytic functions of the variable τ, we obtain that

$$\sum_{k=1}^{\infty} \Theta_{0k}(\sigma',\sigma) e^{-\lambda_{0k}\tau} = \sum_{k=1}^{\infty} \Theta_{1k}(\sigma',\sigma) e^{-\lambda_{1k}\tau} \quad \text{on } \Gamma_{\rm in} \times \Gamma_{\rm out} \tag{3.11}$$

for all $\tau \in (0,\infty)$, where $\Theta_{jk}(\sigma',\sigma) := \sum_{i=1}^{m_{jk}} \psi_{jk,i}(\sigma')\,\psi_{jk,i}(\sigma)$. Now, by Lemma 2.2, we know that $\Theta_{jk}(\sigma',\sigma)$ is not identically zero on any relatively open subset of $\Gamma_{\rm in}\times\Gamma_{\rm out}$. So, using the classical results on Dirichlet's series, (3.11) yields that, for all $k \geq 1$,

$$\lambda_{0k} = \lambda_{1k} \tag{3.12}$$

and

$$\Theta_{0k}(\sigma',\sigma) = \Theta_{1k}(\sigma',\sigma) \quad \text{on } \Gamma_{\rm in}\times\Gamma_{\rm out},$$

that is,

$$\sum_{i=1}^{m_{0k}} \psi_{0k,i}(\sigma')\,\psi_{0k,i}(\sigma) = \sum_{i=1}^{m_{1k}} \psi_{1k,i}(\sigma')\,\psi_{1k,i}(\sigma) \quad \text{on } \Gamma_{\rm in}\times\Gamma_{\rm out}. \tag{3.13}$$

The proof of Lemma 3.4 is complete. □

We are now in a position to prove Theorem 1.1.

Proof of Theorem 1.1. We prove that (3.13) implies that, for all $k \geq 1, m_{0k} = m_{1k}$ and, up to an appropriate choice of the eigenfunctions $\varphi_{0k}, \psi_{0k} = \psi_{1k}$ on $\Gamma_{\rm in}\cup\Gamma_{\rm out}$.

For a fixed $k \geq 1$, let us note that, by Lemma 2.1, $\psi_{jk,i}$, for $i = 1,\dots,m_{jk}$ and $j \in \{0,1\}$, are linearly independent in $L^2(\Gamma_{\rm in}\cup\Gamma_{\rm out})$. Now, applying the algebraic Lemma 2.3 with $m := m_{0k}, n := m_{1k}, Z := \Gamma_{\rm in}\cup\Gamma_{\rm out}, X := \Gamma_{\rm in}, Y := \Gamma_{\rm out}$, and F and G, respectively, the vectors

$$F := \begin{pmatrix} \psi_{0k,1|Z} \\ \vdots \\ \psi_{0k,m_{0k}|Z} \end{pmatrix} \quad \text{and} \quad G := \begin{pmatrix} \psi_{1k,1|Z} \\ \vdots \\ \psi_{1k,m_{1k}|Z} \end{pmatrix},$$

we derive that $m_{0k} = m_{1k}$, and there exists an $m\times m$ orthogonal matrix M, where $m := m_{0k} = m_{1k}$, such that

$$F(z) = MG(z) \quad \text{for } z \in \Gamma_{\rm in}\cup\Gamma_{\rm out}. \tag{3.14}$$

We prove now that $\psi_{0k,i} = \psi_{1k,i}$ on $\Gamma_{\rm in}\cup\Gamma_{\rm out}$, for $i = 1,\dots,m$, up to an appropriate choice of the eigenfunctions $\varphi_{0k,i}$. To prove this, let us define the vector

$$\widetilde{\varphi_0} := M^*\varphi_0^*,$$

where M^* is the transpose matrix of M, that is, $M^*_{ir} = M_{ri}$, and φ_0^* is the transpose vector of $\varphi_0 = (\varphi_{0k,1},\dots,\varphi_{0k,m})$. First let us note that

$$(\widetilde{\varphi_{0k,i}} \mid \widetilde{\varphi_{0k,\ell}}) = \delta_{i\ell} \quad \text{for } 1 \leq i,\ell \leq m,$$

where $(\cdot\mid\cdot)$ denotes the scalar product in $L^2_\rho(\Omega)$, and $\delta_{i\ell}$ is the Kronecker's symbol. In fact, $\widetilde{\varphi_{0k,i}} = \sum_{r=1}^m M^*_{ir}\varphi_{0k,r}$, and $\widetilde{\varphi_{0k,\ell}} = \sum_{s=1}^m M^*_{\ell s}\varphi_{0k,s} = \sum_{s=1}^m \varphi_{0k,s}M_{s\ell}$, so

$$\begin{aligned}(\widetilde{\varphi_{0k,i}} \mid \widetilde{\varphi_{0k,\ell}}) &= \sum_{r=1}^m\sum_{s=1}^m M^*_{ir}M_{s\ell}\delta_{rs} \\ &= \sum_{r=1}^m M^*_{ir}M_{r\ell} = \delta_{i\ell},\end{aligned}$$

where the last equality follows since the matrix M is orthogonal.

Now

$$\widetilde{\psi_{0k,i}} = A_0 \nabla \widetilde{\varphi_{0k,i}} \cdot \mathbf{n} = \sum_{\kappa=1}^{N} \sum_{\iota=1}^{N} a_{0,\kappa\iota} \frac{\partial}{\partial x_\iota} \widetilde{\varphi_{0k,i}} \mathbf{n}_\kappa$$

and

$$\frac{\partial}{\partial x_\iota} \widetilde{\varphi_{0k,i}} = \sum_{\ell=1}^{m} M^*_{i\ell} \frac{\partial}{\partial x_\iota} \varphi_{0k,\ell}.$$

So, substituting in $\widetilde{\psi_{0k,i}}$, we obtain that

$$\begin{aligned} \widetilde{\psi_{0k,i}} &= \sum_{\kappa=1}^{N} \sum_{\iota=1}^{N} a_{0,\kappa\iota} \sum_{\ell=1}^{m} M^*_{i\ell} \frac{\partial}{\partial x_\iota} \varphi_{0k,\ell} \mathbf{n}_\kappa \\ &= \sum_{\ell=1}^{m} M^*_{i\ell} \sum_{\kappa=1}^{N} \sum_{\iota=1}^{N} a_{0,\kappa\iota} \frac{\partial}{\partial x_\iota} \varphi_{0k,\ell} \mathbf{n}_\kappa \\ &= \sum_{\ell=1}^{m} M^*_{i\ell} \psi_{0k,\ell}. \end{aligned}$$

Now, by (3.14), we know that $\psi_{0k,\ell} = \sum_{j=1}^{m} M_{\ell j} \psi_{1k,j}$. Then, substituting in the last equality, it follows that

$$\begin{aligned} \widetilde{\psi_{0k,i}} &= \sum_{\ell=1}^{m} M^*_{i\ell} \sum_{j=1}^{m} M_{\ell j} \psi_{1k,j} \\ &= \sum_{j=1}^{m} \psi_{1k,j} \sum_{\ell=1}^{m} M^*_{i\ell} M_{\ell j} = \psi_{1k,i} \end{aligned}$$

on $\Gamma_{\text{in}} \cup \Gamma_{\text{out}}$, for $1 \le i \le m$, where the last equality follows since the matrix M is orthogonal. The proof of Theorem 1.1 is complete. □

We end this section by proving Theorem 1.2, that is, we show that Theorem 1.1 remains valid if hypothesis (1.12) is replaced by equality of the mean values of the fluxes in the interval $[T_0 - \tau_0, T_0]$, i.e., we suppose that

$$\int_{T_0-\tau_0}^{T_0} A_0 \nabla u_0(t) \cdot \mathbf{n}_{|\Gamma_{\text{out}}} dt = \int_{T_0-\tau_0}^{T_0} A_1 \nabla u_1(t) \cdot \mathbf{n}_{|\Gamma_{\text{out}}} dt \quad \text{in } H^{\frac{1}{2}}(\Gamma_{\text{out}}) \tag{3.15}$$

for all $f \in C([0,T]; H^{\frac{3}{2}}(\Gamma_{\text{in}}))$ such that the support $\text{supp}(f(t,\cdot)) \subset \Gamma_{\text{in}}$ for $t \in [0,T]$.

Proof of Theorem 1.2. First of all, using Lemma 3.3, we can always reduce to the case where in (1.10) the initial data $u_{0j} \equiv 0$ in Ω and the boundary data $\varphi_j \equiv 0$ on $(0,T) \times \partial\Omega \backslash \Gamma_{\text{in}}$. Now, choosing the Dirichlet data f in such a way that $f \equiv 0$ on $[T_0 - \varepsilon_0, T_0] \times \Gamma_{\text{in}}$, where ε_0 is such that $\tau_0 < \varepsilon_0 < T_0$, we write the solutions u_j in the Fourier expansion (3.2), i.e.,

$$u_j(t) = -\sum_{k=1}^{\infty} \int_0^{T_0-\varepsilon_0} \int_{\Gamma_{\text{in}}} \psi_{jk}(\sigma') e^{-\lambda_{jk}(t-\tau)} f(\tau,\sigma') d\sigma' d\tau \varphi_{jk} \quad \text{in } L^2_\rho(\Omega)$$

for $t \in [T_0 - \tau_0, T_0]$. In particular, by Lemma 3.2, we derive that

$$A_j(\sigma)\nabla u_j(t,\sigma)\cdot \mathbf{n}(\sigma) = \int_0^{T_0-\varepsilon_0} \int_{\Gamma_{\rm in}} \sum_{k=1}^{\infty} \Psi_{jk}(\sigma',\sigma,t-\tau) f(\tau,\sigma') d\sigma' d\tau \tag{3.16}$$

for $t \in [T_0 - \tau_0, T_0]$ and a.e. $\sigma \in \partial\Omega$, where Ψ_{jk} are defined in (3.9). Now, (3.15) and the change of variable $t - \tau = s$ in the right-hand side of (3.16) imply that

$$\int_{T_0-\tau_0}^{T_0} \int_{t-T_0+\varepsilon_0}^{t} \int_{\Gamma_{\rm in}} \sum_{k=1}^{\infty} (\Psi_{0k}(\sigma',\sigma,s) - \Psi_{1k}(\sigma',\sigma,s)) f(t-s,\sigma') d\sigma' ds dt = 0. \tag{3.17}$$

We may assume that the input $f(s,\sigma') \equiv 0$ for $s \notin [T' - \varepsilon', T' + \varepsilon']$ and $\sigma' \in \Gamma_{\rm in}$, where $T' \in (\frac{T_0-\tau_0}{2}, \frac{T_0}{2})$ is a fixed time, and $0 < \varepsilon' < T'$. Then (3.17) becomes

$$\int_{T_0-\tau_0}^{T_0} \int_{t-T'-\varepsilon'}^{t-T'+\varepsilon'} \int_{\Gamma_{\rm in}} \sum_{k=1}^{\infty} (\Psi_{0k}(\sigma',\sigma,s) - \Psi_{1k}(\sigma',\sigma,s)) f(t-s,\sigma') d\sigma' ds dt = 0$$

for all such functions f. Hence it follows that

$$\sum_{k=1}^{\infty} \Psi_{0k}(\sigma',\sigma,s) = \sum_{k=1}^{\infty} \Psi_{1k}(\sigma',\sigma,s) \quad \text{on } \Gamma_{\rm in} \times \Gamma_{\rm out}$$

for all $s \in [T' - \varepsilon', T' + \varepsilon']$. Therefore, by the unique continuation principle for analytic functions of the variable s, we obtain that

$$\sum_{k=1}^{\infty} \Theta_{0k}(\sigma',\sigma) e^{-\lambda_{0k} s} = \sum_{k=1}^{\infty} \Theta_{1k}(\sigma',\sigma) e^{-\lambda_{1k} s} \quad \text{on } \Gamma_{\rm in} \times \Gamma_{\rm out} \tag{3.18}$$

for all $s \in (0,\infty)$, where $\Theta_{jk}(\sigma',\sigma) := \sum_{i=1}^{m_{jk}} \psi_{jk,i}(\sigma') \psi_{jk,i}(\sigma)$. Now, using the classical results on Dirichlet series, (3.18) yields that, for all $k \geq 1$,

$$\lambda_{0k} = \lambda_{1k}$$

and

$$\sum_{i=1}^{m_{0k}} \psi_{0k,i}(\sigma') \psi_{0k,i}(\sigma) = \sum_{i=1}^{m_{1k}} \psi_{1k,i}(\sigma') \psi_{1k,i}(\sigma) \quad \text{on } \Gamma_{\rm in} \times \Gamma_{\rm out}.$$

Finally, following the proof of Theorem 1.1, we derive that, for all $k \geq 1, \lambda_{0k} = \lambda_{1k}$ and $\psi_{0k} = \psi_{1k}$ on $\Gamma_{\rm in} \cup \Gamma_{\rm out}$.

The proof of Theorem 1.2 is complete. □

4. Recovering coefficients via BSD. In this section we prove the result stated in Theorem 1.5.

In the last decade many authors have devoted considerable attention to the problem of identifiability of the coefficients in elliptic equations; see, for example, Calderón [6]; Kohn and Vogelius [16, 17]; Sylvester and Uhlmann [26]; Nachman, Sylvester, and Uhlmann [22]; Nachman [23, 24]; Isakov [13]; Alessandrini [2]; Chanillo [9]; Brown [4]; Brown and Uhlmann [5]. More precisely, let us recall the following uniqueness result

(for the proof see Isakov [15, Corollary 5.4.2, p. 119], for $N = 2$, and, for example, Sylvester and Uhlmann [26] for $N \geq 3$; see also the survey paper by Uhlmann [27]).

THEOREM 4.1. *Let $N \geq 2$ be an integer, Ω a bounded domain in $\mathbb{R}^N$ of class $C^{0,1}$, and, for $j \in \{0,1\}$, let (a_j, q_j) be two pairs of functions satisfying conditions (1.14)–(1.18). Let w_j solve*

$$\left\{ \begin{array}{rcl} -\mathrm{div}(a_j \nabla w_j) + q_j w_j &=& 0 \quad \text{in } \Omega, \\ w_j &=& \varphi \quad \text{on } \partial\Omega. \end{array} \right. \tag{4.1}$$

Suppose that, when $N = 2$, either $q_0 = q_1 \equiv 0$ in Ω or $a_0 = a_1 \equiv 1$ in $\overline{\Omega}$ while, when $N \geq 3$, either $q := q_0 = q_1$ or $a := a_0 = a_1$ and

$$\gamma_0(\varphi) = \gamma_1(\varphi) \quad \text{in } H^{-\frac{1}{2}}(\partial\Omega)$$

for all $\varphi \in H^{\frac{1}{2}}(\partial\Omega)$, where $\gamma_j(\varphi) := a_j \frac{\partial}{\partial \mathbf{n}} w_{j|\partial\Omega}$ denote the fluxes of w_j on $\partial\Omega$. Then one has that $a_0 = a_1$ in $\overline{\Omega}$ and $q_0 = q_1$ in Ω.

In 1946 Borg [3] and Levinson [19] asked the following question: does knowledge of the eigenvalues $(\lambda_k)_{k=1}^{\infty}$ of the Sturm–Liouville problem

$$\left\{ \begin{array}{rcll} -v_k'' + q v_k &=& \mu_k v_k & \text{in } (0,\ell), \\ v_k(0) &=& 0, & \\ v_k(\ell) &=& 0 & \end{array} \right.$$

determine q uniquely? It is clear that if $q_0(x) := q(x)$ and $q_1(x) := q(\ell - x)$, the operators A_j, defined by $A_j u := -u'' + q_j u$, have the same eigenvalues: therefore, the spectrum alone in general is not sufficient to determine the potential q uniquely. Later Gel'fand and Levitan [11] proved that knowledge of the eigenvalues $(\lambda_k)_{k=1}^{\infty}$ and of the normalizing constants

$$c_k := \int_0^{\ell} |v_k|^2 \, dx,$$

by supposing $v_k'(0) = 1$, determines q uniquely. In Theorem 1.5 we consider a similar problem in the N-dimensional setting. More precisely, we give a simplified proof of an N-dimensional Borg–Levinson result of Nachman, Sylvester, and Uhlmann [22]. To be in a position to prove Theorem 1.5 we need the following two auxiliary lemmas.

LEMMA 4.2. *Under the assumptions of Theorem 1.5, for $j \in \{0,1\}$, let $w_{j\mu}, \mu \geq 0$, solve*

$$\left\{ \begin{array}{rcl} -\mathrm{div}(a \nabla w_{j\mu}) + (q_j + \mu\rho) w_{j\mu} &=& 0 \quad \text{in } \Omega, \\ w_{j\mu} &=& \varphi \quad \text{on } \partial\Omega. \end{array} \right. \tag{4.2}$$

Then one has that

$$\|w_{j\mu}\|_{L^2_\rho(\Omega)} \to 0 \quad \text{as } \mu \to +\infty.$$

Proof of Lemma 4.2. Since $(\varphi_{jk})_{k=1}^{\infty}$ defined in (1.21) are a Hilbert basis in $L^2_\rho(\Omega)$, we can write $w_{j\mu}$ in the following Fourier expansion:

$$w_{j\mu} = \sum_{k=1}^{\infty} (w_{j\mu} \mid \varphi_{jk}) \varphi_{jk} \quad \text{in } L^2_\rho(\Omega), \tag{4.3}$$

where $(\cdot \mid \cdot)$ denotes the scalar product in $L^2_\rho(\Omega)$. After multiplying (4.2) by φ_{jk}, a direct calculation gives

$$(w_{j\mu} \mid \varphi_{jk}) = \frac{\alpha_{jk}}{\lambda_{jk}+\mu}, \tag{4.4}$$

where $\alpha_{jk} := -\int_{\partial\Omega} \psi_{jk}\varphi d\sigma$ (recall that $\psi_{jk} := a\frac{\partial}{\partial \mathbf{n}}\varphi_{jk|\partial\Omega}$). So, by (4.3) and (4.4), it is easy to verify that

$$\|w_{j\mu}\|_{L^2_\rho(\Omega)} \to 0 \quad \text{as } \mu \to +\infty.$$

The proof is complete. □

LEMMA 4.3. *Under the assumptions of Lemma* 4.2, *let us define* $w_j := w_{j0}$. *Then, for all* $\mu \geq 0$, *one has*

$$\gamma_0(\varphi) - \gamma_1(\varphi) = \gamma_{0\mu}(\varphi) - \gamma_{1\mu}(\varphi) \quad \textit{in } H^{-\frac{1}{2}}(\partial\Omega)$$

for all $\varphi \in H^{\frac{1}{2}}(\partial\Omega)$.

(Recall that $\gamma_j(\varphi) := a\frac{\partial}{\partial \mathbf{n}} w_{j|\partial\Omega}$ and $\gamma_{j\mu}(\varphi) := a\frac{\partial}{\partial \mathbf{n}} w_{j\mu|\partial\Omega}$.)

Proof of Lemma 4.3. Let $w_{j\mu}$ be solutions of (4.2) for $\varphi \in H^{\frac{3}{2}}(\partial\Omega)$. Putting $z_{j\mu} := w_{j\mu} - w_j$, then $z_{j\mu}$ solve

$$\left\{ \begin{array}{rcll} -\mathrm{div}(a\nabla z_{j\mu}) + q_j z_{j\mu} + \mu\rho z_{j\mu} & = & -(\mu-\mu_0)\rho w_j & \text{in } \Omega, \\ z_{j\mu} & = & 0 & \text{on } \partial\Omega \end{array} \right. \tag{4.5}$$

and

$$z_{j\mu} = \sum_{k=1}^{\infty} (z_{j\mu} \mid \varphi_{jk})\varphi_{jk} \quad \text{in } L^2_\rho(\Omega), \tag{4.6}$$

where $(\cdot \mid \cdot)$ denotes the scalar product in $L^2_\rho(\Omega)$. After multiplying (4.5) by φ_{jk}, a direct calculation gives

$$(z_{j\mu} \mid \varphi_{jk}) = -\frac{(\mu-\mu_0)\alpha_k}{(\lambda_k+\mu)^2},$$

where $\alpha_k := -\int_{\partial\Omega} \psi_k\varphi d\sigma, \psi_k := \psi_{0k} = \psi_{1k}, \psi_{jk} := a\frac{\partial}{\partial \mathbf{n}}\varphi_{jk|\partial\Omega}$, and $\lambda_k := \lambda_{0k} = \lambda_{1k}$. Moreover, one can verify that the series (4.6) converge to $z_{j\mu}$ in $H^2(\Omega)$. Now, since the trace operator $\gamma : u \to \frac{\partial}{\partial \mathbf{n}} u_{|\partial\Omega}$ is continuous from $H^2(\Omega)$ to $H^{\frac{1}{2}}(\partial\Omega)$, it follows that

$$a\frac{\partial}{\partial \mathbf{n}} z_{j\mu|\partial\Omega} = -\sum_{k=1}^{\infty} \frac{(\mu-\mu_0)\alpha_k}{(\lambda_k+\mu)^2}\psi_{jk} = -\sum_{k=1}^{\infty} \frac{(\mu-\mu_0)\alpha_k}{(\lambda_k+\mu)^2}\psi_k \quad \text{in } H^{\frac{1}{2}}(\partial\Omega).$$

So, in the sense of $H^{\frac{1}{2}}(\partial\Omega)$, one has

$$a\frac{\partial}{\partial \mathbf{n}} z_{0\mu|\partial\Omega} - a\frac{\partial}{\partial \mathbf{n}} z_{1\mu|\partial\Omega} = 0,$$

that is,

$$\gamma_0(\varphi) - \gamma_1(\varphi) = \gamma_{0\mu}(\varphi) - \gamma_{1\mu}(\varphi) \quad \text{in } H^{\frac{1}{2}}(\partial\Omega) \tag{4.7}$$

for all $\varphi \in H^{\frac{3}{2}}(\partial\Omega)$ and for all $\mu \geq 0$. Finally, by a density argument, (4.7) holds in $H^{-\frac{1}{2}}(\partial\Omega)$ for all $\varphi \in H^{\frac{1}{2}}(\partial\Omega)$.

The proof is complete. □

Now we can prove Theorem 1.5.

Proof of Theorem 1.5. Let $w_{j\mu}$ solve (4.2) for $\varphi \in H^{\frac{3}{2}}(\partial\Omega), \mu \geq 0$. Multiplying (4.2), for $j = 0$, by $w_{1\mu}$ and integrating by parts over Ω we obtain

$$\int_\Omega a\nabla w_{0\mu} \cdot \nabla w_{1\mu} dx - \int_{\partial\Omega} a\frac{\partial}{\partial \mathbf{n}} w_{0\mu}\varphi d\sigma + \int_\Omega q_0 w_{0\mu} w_{1\mu} dx = -\mu \int_\Omega \rho w_{0\mu} w_{1\mu} dx. \tag{4.8}$$

Similarly, multiplying (4.2), for $j = 1$, by $w_{0\mu}$ and integrating by parts over Ω we obtain

$$\int_\Omega a\nabla w_{0\mu} \cdot \nabla w_{1\mu} dx - \int_{\partial\Omega} a\frac{\partial}{\partial \mathbf{n}} w_{1\mu}\varphi d\sigma + \int_\Omega q_1 w_{0\mu} w_{1\mu} dx = -\mu \int_\Omega \rho w_{0\mu} w_{1\mu} dx. \tag{4.9}$$

Then, subtracting (4.8) from (4.9), it follows that

$$\int_{\partial\Omega} (\gamma_{0\mu}(\varphi) - \gamma_{1\mu}(\varphi))\varphi d\sigma + \int_\Omega (q_1 - q_0) w_{0\mu} w_{1\mu} dx = 0 \tag{4.10}$$

for all $\mu \geq 0$. Thus, letting $\mu \to +\infty$ in (4.10), recalling that $\|w_{j\mu}\|_{L^2_\rho(\Omega)} \to 0$ as $\mu \to +\infty$ and that $\gamma_0(\varphi) - \gamma_1(\varphi) = \gamma_{0\mu}(\varphi) - \gamma_{1\mu}(\varphi)$, for all $\mu \geq 0$, we derive that

$$\int_{\partial\Omega} (\gamma_0(\varphi) - \gamma_1(\varphi))\varphi d\sigma = 0 \tag{4.11}$$

for all $\varphi \in H^{\frac{3}{2}}(\partial\Omega)$. In particular, by a density argument, (4.11) holds for all $\varphi \in H^{\frac{1}{2}}(\partial\Omega)$, which implies

$$\gamma_0(\varphi) - \gamma_1(\varphi) = 0 \quad \text{in } H^{-\frac{1}{2}}(\partial\Omega)$$

for all $\varphi \in H^{\frac{1}{2}}(\partial\Omega)$. Finally, by Theorem 4.1, it follows that $q_0 = q_1$ in Ω.
The proof of Theorem 1.5 is complete. □

5. Uniqueness results for some classes of heat equations. In this section we prove the injectivity of the operator Φ defined in (1.7) in the following three cases:

(i) given $q(x)$, and $A(x)$ of the form $A(x) = a(x)I_N$, where $a(x)$ is a scalar-valued function and I_N is the identity matrix, we identify $\rho(x)$ and $a(x)$ by supposing that $\Gamma_{\text{in}} = \Gamma_{\text{out}} = \partial\Omega$;

(ii) given $A(x) = a(x)I_N$, we identify $\rho(x)$ and $q(x)$ by supposing that $\Gamma_{\text{in}} = \Gamma_{\text{out}} = \partial\Omega$;

(iii) given $\rho(x)$ and $A(x) = a(x)I_N$, we identify $q(x)$ by supposing that $\Gamma_{\text{in}} \cup \Gamma_{\text{out}} = \partial\Omega$.

We begin by proving case (i), that is, Theorem 1.3: given q, we prove the identifiability of ρ and a.

Proof of Theorem 1.3. First let us note that, without loss of generality, we can suppose that in (1.10) the initial data $u_{0j} \equiv 0$ in $\Omega, j \in \{0, 1\}$ (see Lemma 3.3). As usual we denote by $(\lambda_{jk})_{k=1}^\infty$ and $(\varphi_{jk})_{k=1}^\infty$, respectively, the eigenvalues and the corresponding eigenfunctions of the underlying elliptic operators with Dirichlet boundary conditions, that is,

$$\left\{ \begin{array}{rcll} -\text{div}(a_j \nabla \varphi_{jk}) + q\varphi_{jk} & = & \lambda_{jk}\rho_j\varphi_{jk} & \text{in } \Omega, \\ \varphi_{jk} & = & 0 & \text{on } \partial\Omega, \\ \int_\Omega |\varphi_{jk}|^2 \rho_j dx & = & 1. & \end{array} \right. \tag{5.1}$$

By Theorem 1.1 we know that the boundary spectral data $\mathrm{BSD}(\rho_j, a_j, q)$ coincide, that is, for all $k \geq 1$,

$$\lambda_{0k} = \lambda_{1k} =: \lambda_k \quad \text{and} \quad \psi_{0k} = \psi_{1k} =: \psi_k \quad \text{on } \partial\Omega. \tag{5.2}$$

(We recall that $\psi_{jk} := a_j \frac{\partial}{\partial \mathbf{n}} \varphi_{jk|\partial\Omega}$.) Set $u_j(t,x) = v_j(t,x) + w_j(x)$, where v_j solve

$$\begin{cases} \rho_j(x)\partial_t v_j - \mathrm{div}(a_j(x)\nabla v_j) + q(x)v_j &= 0 & \text{in } (0,T)\times\Omega, \\ v_j(0) &= -w_j & \text{in } \Omega, \\ v_j(t,\sigma) &= 0 & \text{on } (0,T)\times\partial\Omega, \end{cases}$$

and w_j solve (4.1) for $\varphi \in H^{\frac{3}{2}}(\partial\Omega)$. Since $(\varphi_{jk})_{k=1}^{\infty}$ are a Hilbert basis in $L^2_\rho(\Omega)$, we can write $v_j(t)$ in the following Fourier expansion:

$$v_j(t) = \sum_{k=1}^{\infty} \alpha_{jk}(t)\varphi_{jk} \quad \text{in } L^2_\rho(\Omega), \tag{5.3}$$

where the coefficients $\alpha_{jk}(t) := (v_j(t) \mid \varphi_{jk})$ solve

$$\begin{cases} \frac{d}{dt}\alpha_{jk}(t) + \lambda_{jk}\alpha_{jk}(t) &= 0 & \text{in } (0,T), \\ \alpha_{jk}(0) &= -(w_j \mid \varphi_{jk}), \end{cases}$$

and $(\cdot \mid \cdot)$ denotes the scalar product in $L^2_\rho(\Omega)$. Then

$$\alpha_{jk}(t) = -(w_j \mid \varphi_{jk})e^{-\lambda_{jk}t}.$$

After multiplying (4.1) by φ_{jk}, a simple calculation gives

$$(w_j \mid \varphi_{jk}) = -\frac{1}{\lambda_{jk}} \int_{\partial\Omega} \psi_{jk}\varphi d\sigma.$$

Hence, from (5.2), we deduce that $\alpha_{0k}(t) = \alpha_{1k}(t) =: \alpha_k(t)$ on $[0,T]$ for all $k \geq 1$.

Moreover, one can verify that the series (5.3), at $t = T_0$, converge to $v_j(T_0)$ in $H^2(\Omega)$. Thus, since the trace operator $\gamma : u \to \frac{\partial}{\partial \mathbf{n}} u_{|\partial\Omega}$ is continuous from $H^2(\Omega)$ to $H^{\frac{1}{2}}(\partial\Omega)$, it follows that

$$a_j \frac{\partial}{\partial \mathbf{n}} v_j(T_0)_{|\partial\Omega} = \sum_{k=1}^{\infty} \alpha_k(T_0)\psi_{jk} = \sum_{k=1}^{\infty} \alpha_k(T_0)\psi_k \quad \text{in } H^{\frac{1}{2}}(\partial\Omega),$$

and therefore

$$a_j \frac{\partial}{\partial \mathbf{n}} u_j(T_0)_{|\partial\Omega} = \sum_{k=1}^{\infty} \alpha_k(T_0)\psi_k + a_j \frac{\partial}{\partial \mathbf{n}} w_{j|\partial\Omega} \quad \text{in } H^{\frac{1}{2}}(\partial\Omega). \tag{5.4}$$

Now, recalling that $a_0 \frac{\partial}{\partial \mathbf{n}} u_0(T_0)_{|\partial\Omega} = a_1 \frac{\partial}{\partial \mathbf{n}} u_1(T_0)_{|\partial\Omega}$, from (5.4) we derive that

$$a_0 \frac{\partial}{\partial \mathbf{n}} w_{0|\partial\Omega} = a_1 \frac{\partial}{\partial \mathbf{n}} w_{1|\partial\Omega} \quad \text{in } H^{\frac{1}{2}}(\partial\Omega) \tag{5.5}$$

for all $\varphi \in H^{\frac{3}{2}}(\partial\Omega)$. In particular, by a density argument, (5.5) holds in $H^{-\frac{1}{2}}(\partial\Omega)$ for all $\varphi \in H^{\frac{1}{2}}(\partial\Omega)$. Finally, by Theorem 4.1, it follows that $a_0 = a_1$ in $\overline{\Omega}$.

Now we denote by $a(x) := a_0(x) = a_1(x)$. Set $u_j(t,x) = \widetilde{v_j}(t,x) + \widetilde{w_j}(x)$, where $\widetilde{v_j}$ solve

$$\left\{\begin{array}{rcll} \rho_j(x)\partial_t\widetilde{v_j} - \mathrm{div}(a(x)\nabla\widetilde{v_j}) + q(x)\widetilde{v_j} &=& -\rho_j(x)\widetilde{w_j} & \text{in } (0,T)\times\Omega, \\ \widetilde{v_j}(0) &=& -\widetilde{w_j} & \text{in } \Omega, \\ \widetilde{v_j}(t,\sigma) &=& 0 & \text{on } (0,T)\times\partial\Omega, \end{array}\right.$$

and $\widetilde{w_j}$ solve

$$\left\{\begin{array}{rcll} -\mathrm{div}(a\nabla\widetilde{w_j}) + (q+\rho_j)\widetilde{w_j} &=& 0 & \text{in } \Omega, \\ \widetilde{w_j} &=& \varphi & \text{on } \partial\Omega. \end{array}\right.$$

By following the above arguments, again we derive that

$$a\frac{\partial}{\partial\mathbf{n}}\widetilde{w_0}_{|\partial\Omega} = a\frac{\partial}{\partial\mathbf{n}}\widetilde{w_1}_{|\partial\Omega} \quad \text{in } H^{-\frac{1}{2}}(\partial\Omega)$$

for all $\varphi \in H^{\frac{1}{2}}(\partial\Omega)$. Then Theorem 4.1 implies that $\rho_0 = \rho_1$ in Ω.

The proof of Theorem 1.3 is complete. □

Now we prove case (ii), that is, Theorem 1.4: given a, we prove the identifiability of ρ and q.

Proof of Theorem 1.4. Repeating arguments, with the obvious changes, in the proof of Theorem 1.3 one obtains, in a first step, that $q_0 = q_1$ in Ω, and in a second step, that $\rho_0 = \rho_1$ in Ω. □

As a direct consequence of Theorems 1.1 and 1.5, we can prove case (iii), that is, Theorem 1.6: given ρ and a, we prove the identifiability of q by supposing that $\Gamma_{\mathrm{in}} \cup \Gamma_{\mathrm{out}} = \partial\Omega$.

Proof of Theorem 1.6. As usual, without loss of generality, we can suppose that in (1.10) the initial data $u_{0j} \equiv 0$ in Ω and the boundary data $\varphi_j \equiv 0$ on $(0,T)\times\partial\Omega\backslash\Gamma_{\mathrm{in}}$. We denote by $(\lambda_{jk})_{k=1}^{\infty}$ and $(\varphi_{jk})_{k=1}^{\infty}$, respectively, the eigenvalues and the corresponding eigenfunctions of the underlying elliptic operators with Dirichlet boundary conditions, that is,

$$\left\{\begin{array}{rcll} -\mathrm{div}(a\nabla\varphi_{jk}) + q_j\varphi_{jk} &=& \lambda_{jk}\rho\varphi_{jk} & \text{in } \Omega, \\ \varphi_{jk} &=& 0 & \text{on } \partial\Omega, \\ \int_\Omega |\varphi_{jk}|^2 \rho dx &=& 1. & \end{array}\right.$$

By Theorem 1.1 we know that the boundary spectral data $\mathrm{BSD}(\rho, a, q_j)$ coincide, that is, for all $k \geq 1$,

$$\lambda_{0k} = \lambda_{1k} \quad \text{and} \quad \psi_{0k} = \psi_{1k} \quad \text{on} \quad \partial\Omega.$$

So, by Theorem 1.5, it follows that $q_0 = q_1$ in Ω.

The proof is complete. □

6. The one dimensional case. In this section we investigate the problem of determining the coefficients $\rho(x), a(x), q(x)$ in the one dimensional heat equation

$$\left\{\begin{array}{rcll} \rho(x)\partial_t u - \partial_x(a(x)\partial_x u) + q(x)u &=& 0 & \text{in } (0,T)\times(0,\ell), \\ u(0) &=& u_0 & \text{in } (0,\ell), \\ u(t,0) &=& f(t) & \text{in } (0,T), \\ u(t,\ell) &=& g(t) & \text{in } (0,T) \end{array}\right. \tag{6.1}$$

from the additional measurement of the thermal flux $a(x_*)\partial_x u(T_0, x_*)$ at a given time $T_0 \in (0, T]$, at one end-point x_* of the interval $[0, \ell]$, when all input $f \in C[0, T]$ are assigned in (6.1) (the temperature $u(t, \ell) = g(t)$ at the other end-point $x = \ell$ is given arbitrarily). We study two different cases. In the first one we prove that knowledge of the additional data $a(0)\partial_x u(T_0, 0)$ at the end-point $x_* = 0$, where the input f is assigned, uniquely determines *one* of the coefficients ρ, a, q. In the second case we prove that knowledge of the additional data $a(\ell)\partial_x u(T_0, \ell)$ at the other end-point $x_* = \ell$ (where the temperature $u(t, \ell)$ is a priori unknown) in general does not determine uniquely ρ, a, q, except for symmetric coefficients, in which case uniqueness holds.

We make the following regularity assumptions on the coefficients ρ, a, q:

$$\rho \in C^{1,1}[0, \ell] \text{ and } \rho \geq \beta \text{ for some constant } \beta > 0; \tag{6.2}$$

$$a \in C^{1,1}[0, \ell] \text{ and } a \geq \alpha \text{ for some constant } \alpha > 0; \tag{6.3}$$

$$q \in L^p(0, \ell) \text{ for some } p > 1. \tag{6.4}$$

In this section we prove the following uniqueness result.

THEOREM 6.1. *Let $N = 1, \ell < +\infty$ and, for $j \in \{0, 1\}$, let (ρ_j, a_j, q_j) be two sets of functions satisfying conditions* (6.2)–(6.4). *For some fixed $u_{0j} \in C[0, \ell]$ and $g_j \in C[0, T]$, let u_j be solutions of* (6.1) *when $(\rho, a, q) := (\rho_j, a_j, q_j), u_0 := u_{0j}$ and $g := g_j$. We denote by*

$$\Lambda_j(f) := a_j(0)\partial_x u_j(T_0, 0)$$

the thermal fluxes measured at a given time $T_0 \in (0, T]$ at the end-point $x_* = 0$. Suppose that one has

$$\Lambda_0(f) = \Lambda_1(f)$$

for all $f \in C[0, T]$ such that $f \equiv 0$ in $[T_0 - \varepsilon_0, T_0]$, where ε_0 is such that $0 < \varepsilon_0 < T_0$. Then, given $\rho := \rho_0 = \rho_1$ and $q := q_0 = q_1$, one has that $a_0 = a_1$ in $[0, \ell]$.

REMARK 6.2. We observe that one can prove that either, given $\rho := \rho_0 = \rho_1$ and $a := a_0 = a_1$, that $q_0 = q_1$ in $(0, \ell)$, or given $a := a_0 = a_1$ and $q := q_0 = q_1$, that $\rho_0 = \rho_1$ in $[0, \ell]$. □

Proof of Theorem 6.1. First of all, without loss of generality, we can suppose that the initial data $u_{0j} \equiv 0$ in $(0, \ell)$ and the boundary data $g_j \equiv 0$ in $(0, T)$ (see Lemma 3.3). Now, let us denote by $(\lambda_{jk})_{k=1}^{\infty}$ and $(\varphi_{jk})_{k=1}^{\infty}$, respectively, the eigenvalues and the corresponding eigenfunctions of the following problems (with Dirichlet boundary conditions):

$$\begin{cases} -(a_j\varphi'_{jk})' + q_j\varphi_{jk} &= \lambda_{jk}\rho_j\varphi_{jk} \quad \text{in } (0, \ell), \\ \varphi_{jk}(0) = \varphi_{jk}(\ell) &= 0, \\ \int_0^\ell |\varphi_{jk}|^2 \rho_j dx &= 1. \end{cases} \tag{6.5}$$

By Theorem 1.1 we know that the boundary spectral data $\mathrm{BSD}(\rho_j, a_j, q_j) := (\lambda_{jk}, a_j(0)\varphi'_{jk}(0))$ coincide, i.e., for all $k \geq 1$, one has that

$$\lambda_{0k} = \lambda_{1k} =: \lambda_k \tag{6.6}$$

and

$$a_0(0)\varphi'_{0k}(0) = a_1(0)\varphi'_{1k}(0). \tag{6.7}$$

Now, supposing in (6.5) that $\rho := \rho_0 = \rho_1$ in $[0,\ell]$ and $q := q_0 = q_1$ in $(0,\ell)$, we prove that (6.6) and (6.7) imply that $a_0 = a_1$ in $[0,\ell]$. Since the proof of this result is reasonably well known, for the reader's convenience we give only an outline of it (one may refer to McLaughlin [21], Gladwell [12], Chadan, Colton, Päivärinta, and Rundell [8]. We thank one of the referees for having suggested these references). First, recalling the asymptotic behavior of the eigenvalues of (6.5) (see, for example, Courant and Hilbert [10, p. 414]), and (6.6), one can derive that

$$\int_0^\ell \left(\frac{\rho}{a_0}\right)^{\frac{1}{2}} ds = \int_0^\ell \left(\frac{\rho}{a_1}\right)^{\frac{1}{2}} ds =: \ell_*.$$

Then, we consider the so-called Liouville transform of the eigenfunctions φ_{jk} (see, for example, Courant and Hilbert [10, p. 292]), i.e., setting

$$y := \alpha_j(x) = \int_0^x \left(\frac{\rho}{a_j}\right)^{\frac{1}{2}} ds, \tag{6.8}$$

we define

$$v_{jk}(y) := (\widetilde{a_j}(y)\widetilde{\rho}(y))^{\frac{1}{4}}\widetilde{\varphi_{jk}}(y), \tag{6.9}$$

where $\widetilde{a_j}(y) := a_j(\alpha_j^{-1}(y)), \widetilde{\rho}(y) := \rho(\alpha_j^{-1}(y))$ and $\widetilde{\varphi_{jk}}(y) := \varphi_{jk}(\alpha_j^{-1}(y))$. One can verify that the functions v_{jk} satisfy the following problems (the so-called Sturm–Liouville problem in normal form):

$$\left\{\begin{array}{rcl} v''_{jk} - p_j v_{jk} &=& \lambda_{jk} v_{jk} \quad \text{in } (0,\ell_*), \\ v_{jk}(0) = v_{jk}(\ell_*) &=& 0, \\ \int_0^{\ell^*} |v_{jk}|^2\, dy &=& 1, \end{array}\right. \tag{6.10}$$

where

$$p_j(y) := \frac{f''_j(y)}{f_j(y)} + \frac{\widetilde{q}(y)}{\widetilde{\rho}(y)}, \tag{6.11}$$

$$f_j(y) := (\widetilde{a_j}(y)\widetilde{\rho}(y))^{\frac{1}{4}}, \tag{6.12}$$

and $\widetilde{q}(y) := q(\alpha_j^{-1}(y))$. Recalling the asymptotic behavior of the eigenfunctions of (6.10) (see, for example, Courant and Hilbert [10, p. 338]), and (6.7) it is easy to obtain that

$$a_0(0) = a_1(0)$$

and then, for all $k \geq 1$,

$$v'_{0k}(0) = v'_{1k}(0).$$

Thus, following Gel'fand and Levitan [11], the coefficients p_j in (6.10) can be computed via

$$p_j(y) = \frac{1}{2}\frac{d}{dy}K(y,y),$$

where $K(t,y)$ is the unique solution of the linear integral equation

$$F(t,y)+K(t,y)+\int_0^y K(s,y)F(t,s)ds=0,\quad 0\le t\le y\le \ell_*,$$

where

$$F(t,y)=\frac{1}{\xi_1}\sin(\sqrt{\lambda_1}y)(\sqrt{\lambda_1}t)-\frac{1}{\ell_*}+\sum_{k=2}^{\infty}\left[\frac{\sin(\sqrt{\lambda_k}y)\sin(\sqrt{\lambda_k}t)}{\xi_k}-\frac{2}{\ell_*}\sin\left(\frac{k\pi}{\ell_*}y\right)\sin\left(\frac{k\pi}{\ell_*}t\right)\right]$$

and $\xi_k:=v_k'(0)^{-2}$, where $v_k'(0):=v_{0k}'(0)=v_{1k}'(0)$. Thus we derive that $p_0=p_1$ in $[0,\ell]$.

The proof of Theorem 6.1 is completed by going back to the coefficients a_j of (6.5). □

We conclude this section with an example in which we prove that knowledge, for all possible input f assigned in $x=0$, of the additional data $a(\ell)\partial_x u(T_0,\ell)$ at the end-point $x_*=\ell$ of the interval $[0,\ell]$, where the datum g is supposed to be identically zero, in general does not determine uniquely the coefficients ρ,a,q. This example is contained in the following proposition.

PROPOSITION 6.3. *Under the assumptions of Theorem* 6.1, *let* $\rho:=\rho_0=\rho_1, q:=q_0=q_1$ *be given, and let* $\rho,q\in C[0,\ell]$ *be symmetric functions, that is,* $\rho(x)=\rho(\ell-x)$ *and* $q(x)=q(\ell-x)$. *For* $a\in C^1[0,\ell]$, *let* $a_0(x):=a(x)$, *and let* $a_1(x):=a(\ell-x)$. *Suppose that* $u_j, j\in\{0,1\}$, *solve* (6.1) *when* $(\rho,a,q):=(\rho,a_j,q), u_0\equiv 0$ *in* $(0,\ell)$ *and* $g\equiv 0$ *in* $(0,T)$. *Then, for a given* $T_0\in(0,T]$, *and for all* $f\in C(0,T)$ *such that* $f\equiv 0$ *in* $[T_0-\varepsilon_0,T_0]$, *where* $0<\varepsilon_0<T_0$, *one has*

$$a_0(\ell)\partial_x u_0(T_0,\ell)=a_1(\ell)\partial_x u_1(T_0,\ell).$$

Proof of Proposition 6.3. First let us note that, since the coefficients ρ,q are symmetric and $a_0(x)=a_1(\ell-x)$, it follows that the eigenvalues and the eigenfunctions of problems (6.5) verify the following identities:

$$\lambda_{0k}=\lambda_{1k}\quad\text{and}\quad\varphi_{1k}(x)=\varphi_{0k}(\ell-x)\tag{6.13}$$

for all $k\ge 1$. By Lemma 3.2 we know that

$$a_j(\ell)\partial_x u_j(T_0,\ell)=-\sum_{k=1}^{\infty}\int_0^{T_0-\varepsilon_0}e^{-\lambda_{jk}(T_0-\tau)}a_j(0)a_j(l)\varphi_{jk}'(0)\varphi_{jk}'(\ell)f(\tau)d\tau.$$

Thus (6.13) yields that $a_0(\ell)\partial_x u_0(T_0,\ell)=a_1(\ell)\partial_x u_1(T_0,\ell)$. □

REMARK 6.4. Under the assumptions of Theorem 6.1, let $\rho:=\rho_0=\rho_1, q:=q_0=q_1$ be given, and, for $j\in\{0,1\}$, let ρ,a_j,q be symmetric functions, that is, $\rho(x)=\rho(\ell-x), a_j(x)=a_j(\ell-x), q(x)=q(\ell-x)$. Suppose that the fluxes of the solutions u_j of (6.1), when $(\rho,a,q):=(\rho,a_j,q), u_0:=u_{0j}$, and $g:=g_j$, coincide at the end-point $x_*=\ell$, i.e.,

$$a_0(\ell)\partial_x u_0(T_0,\ell)=a_1(\ell)\partial_x u_1(T_0,\ell)$$

for all $f\in C[0,T]$ such that $f=0$ in $[T_0-\varepsilon_0,T_0]$. Then $a_0=a_1$ in $[0,\ell]$. □

REFERENCES

[1] V. Adolfsson and L. Escauriaza, *$C^{1,\alpha}$ domains and unique continuation at the boundary*, Comm. Pure Appl. Math., 50 (1997), pp. 935–969.

[2] G. Alessandrini, *Singular solutions of elliptic equations and the determination of conductivity by boundary measurements*, J. Differential Equations, 84 (1990), pp. 252–273.

[3] G. Borg, *Eine umkehrung der Sturm-Liouville eigenwertaufgabe*, Acta Math., 78 (1946), pp. 1–96.

[4] R.M. Brown, *Global uniqueness in the impedance-imaging problem for less regular conductivities*, SIAM J. Math. Anal., 27 (1996), pp. 1049–1056.

[5] R.M. Brown and G. Uhlmann, *Uniqueness in the inverse conductivity problem for nonsmooth conductivities in two dimensions*, Comm. Partial Differential Equations, 22 (1997), pp. 1009–1027.

[6] A.P. Calderón, *On an inverse boundary value problem*, Seminar on Numerical Analysis and its Applications to Continuum Physics, Soc. Brasileira de Matemática, Rio de Janeiro, 1980, pp. 65–73.

[7] T. Cazenave and A. Haraux, *An Introduction to Semilinear Evolution Equations*, Oxford Lecture Ser. Math. Appl. 13, Oxford University Press, New York, 1999.

[8] K. Chadan, D. Colton, L. Päivärinta, and W. Rundell, *An Introduction to Inverse Scattering and Inverse Spectral Problems*, SIAM Monogr. Math. Model. Comput., SIAM, Philadelphia, 1997.

[9] S. Chanillo, *A problem in electrical prospection and an n-dimensional Borg-Levinson theorem*, Proc. Amer. Math. Soc., 108 (1990), pp. 761–767.

[10] R. Courant and D. Hilbert, *Methods of Mathematical Physics*, *Vol.* I, Wiley-Interscience, New York, 1989.

[11] I.M. Gel'fand and B.M. Levitan, *On the determination of a differential equation from its spectral function*, Izv. Akad. Nauk. SSSR Ser. Mat., 15 (1951), pp. 309–360; translation in Amer. Math. Soc. Transl. (2), 1 (1955), pp. 253–304.

[12] G.M.L. Gladwell, *Inverse problems in scattering. An introduction*, in Solid Mech. Appl. 23, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1993.

[13] V. Isakov, *On uniqueness of recovery of a discontinuous conductivity coefficient*, Comm. Pure Appl. Math., 41 (1988), pp. 865–877.

[14] V. Isakov, *Uniqueness and stability in multidimensional inverse problems*, Inverse Problems, 9 (1993), pp. 579–621.

[15] V. Isakov, *Inverse Problems for Partial Differential Equations*, in Appl. Math. Sci. 127, Springer, New York, 1998.

[16] R.V. Kohn and M. Vogelius, *Determining conductivity by boundary measurements*, Comm. Pure Appl. Math., 37 (1984), pp. 289–297.

[17] R.V. Kohn and M. Vogelius, *Determining conductivity by boundary measurements, interior results,* II, Comm. Pure Appl. Math., 38 (1985), pp. 643–667.

[18] C. Kravaris and J.H. Seinfeld, *Identifiability of spatially-varying conductivity from point observation as an inverse Sturm-Liouville problem*, SIAM J. Control Optim., 24 (1986), pp. 522–542.

[19] N. Levinson, *The inverse Sturm-Liouville problem*, Mat. Tidsskr. B, 1949, pp. 25–30.

[20] J.-L. Lions and E. Magenes, *Problèmes aux Limites non Homogènes et Applications*, *Vol.* 1, Dunod, Paris, 1968.

[21] J.R. McLaughlin, *Analytical methods for recovering coefficients in differential equations from spectral data*, SIAM Rev., 28 (1986), pp. 53–72.

[22] A.I. Nachman, J. Sylvester, and G. Uhlmann, *An n-dimensional Borg-Levinson theorem*, Comm. Math. Phys., 115 (1988), pp. 595–605.

[23] A.I. Nachman, *Reconstructions from boundary measurements*, Ann. Math., 128 (1988), pp. 531–577.

[24] A.I. Nachman, *Global uniqueness for a two-dimensional inverse boundary value problem*, Ann. Math., 142 (1996), pp. 71–96.

[25] Rakesh and W.W. Symes, *Uniqueness for an inverse problem for the wave equation*, Comm. Partial Differential Equations, 13 (1988), pp. 87–96.

[26] J. Sylvester and G. Uhlmann, *Global uniqueness theorem for an inverse boundary problem*, Ann. Math., 41 (1987), pp. 153–169.

[27] G. Uhlmann, *Inverse boundary value problems and applications*, Astérisque, 207 (1992), pp. 153–169.

SIAM J. MATH. ANAL.
Vol. 32, No. 5, pp. 987–1013

UNIFORM ASYMPTOTIC EXPANSIONS FOR THE REVERSE GENERALIZED BESSEL POLYNOMIALS, AND RELATED FUNCTIONS*

T. M. DUNSTER†

Abstract. The generalized Bessel polynomials $y_n(z;a)$ are generalizations of the well-known modified Bessel functions $K_\nu(z)$. In this paper a study is undertaken of the reverse generalized Bessel polynomials, which are defined by $\theta_n(z;a) = z^n y_n(z^{-1};a)$, and their asymptotic behavior as the degree $n \to \infty$ is comprehensively determined. This is achieved by an application of two general asymptotic theories, due to F. W. J. Olver, to an ordinary differential equation satisfied by $\theta_n(z;a)$: one for the case of complex domains which are free of turning points (yielding Liouville–Green expansions), and the other for complex domains containing a simple turning point (yielding Airy function expansions). The approximations are uniformly valid for z lying in certain unbounded subdomains of the complex plane and are complete with explicit error bounds. Together the domains of validity cover the whole complex plane. Moreover, all the results are uniformly valid for real or complex a, the only restrictions being that $a = O(n)$ as $n \to \infty$ and that a must not be close to the values $-n$ or $-2n$. Two companion solutions to $\theta_n(z;a)$ are defined, which are solutions of the same differential equation and are recessive at certain singularities of the equation. Similar Liouville–Green and Airy function expansions, with accompanying error bounds, are also derived for these functions.

Key words. Bessel polynomials, Liouville–Green expansions, turning point problems

AMS subject classifications. Primary, 33C15; Secondary, 33C10, 34E20

PII. S0036141099359068

1. Introduction.

The generalized Bessel polynomials are defined by

$$y_n(z;a) = \sum_{k=0}^{n} \binom{n}{k} (n+a-1)_k \left(\frac{1}{2}z\right)^k, \tag{1.1}$$

where $(\alpha)_k = \Gamma(\alpha+k)/\Gamma(\alpha)$ is Pochhammer's symbol. In terms of the hypergeometric function they can be represented by

$$y_n(z;a) = {}_2F_0\left(-n, a+n-1; -\frac{1}{2}z\right). \tag{1.2}$$

These polynomials satisfy the linear second order differential equation

$$z^2 \frac{\mathrm{d}^2 y}{\mathrm{d}z^2} + (az+2)\frac{\mathrm{d}y}{\mathrm{d}z} - n(n+a-1)y = 0, \tag{1.3}$$

which has an irregular singularity at $z = 0$ and a regular singularity at infinity. We find it preferable to have the location of these types of singularities interchanged, and therefore, we shall study the so-called reverse generalized Bessel polynomials, which are defined by

$$\theta_n(z;a) = z^n y_n(z^{-1};a). \tag{1.4}$$

*Received by the editors July 16, 1999; accepted for publication August 15, 2000; published electronically January 16, 2001. This research was supported by the National Science Foundation under grant DMS-9970489.

http://www.siam.org/journals/sima/32-5/35906.html

†Department of Mathematical Sciences, San Diego State University, San Diego, CA 92182-7720 (dunster@math.sdsu.edu).

The purpose of this paper is to investigate the asymptotic behavior of the polynomials $\theta_n(z;a)$ as $n \to \infty$. Clearly, from the relationship (1.4), any asymptotic results concerning $\theta_n(z;a)$, or their zeros, trivially lead to corresponding results for the generalized Bessel polynomials $y_n(z;a)$: the domain of validity of any asymptotic results for $\theta_n(z;a)$ must, of course, be transformed under the map $z \to 1/z$ to give the domain of validity of the corresponding asymptotic results for $y_n(z;a)$. However, our asymptotic results for $\theta_n(z;a)$, when taken together, will be valid for all (unbounded) complex values of z, and consequently, the same is true when these results are converted for the generalized Bessel polynomials $y_n(z;a)$.

An early appearance of the generalized Bessel polynomials was in the 1929 papers by Bochner [1] and Romanovsky [10], and they have appeared after that in papers by many other authors. The importance of these polynomials was seen in 1949 by Krall and Frink [7] (who introduced the name Bessel polynomials) in their connection with the wave equation in spherical coordinates, and also in 1949 by Thompson [11] in his study of electrical networks.

To this day, the asymptotic behavior of $y_n(z;a)$ for large degree n and real or complex z has not been fully determined, due to lack of error bounds and restrictive regions of validity. Perhaps the reason for this is that even for real positive a, the problem involves turning points which are neither real nor purely imaginary: the only exception being when $a = 2$ (leading to purely imaginary turning points), in which case well-known asymptotic results for unmodified and modified Bessel functions are applicable (see (1.12) below, [8, Chap. 10, sect. 7], and [8, Chap. 11, sect. 10]). The position of the turning points in the complex plane makes the analysis more complicated, both for integral and for differential equation methods. In the latter method, which we shall use in this paper, the primary difficulty of having complex turning points is that they lead to a more complicated Liouville transformation (details of which are given in section 3 of this paper).

The most recent, and comprehensive, asymptotic work on generalized Bessel polynomials is that of Wong and Zhang [12], who used an integral representation and the method of Chester, Friedman, and Ursel [4] to obtain asymptotic expansions for $y_n(z;a)$ involving Airy functions. In this paper we obtain similar (but not identical) Airy function expansions: our results are more extensive than those of [12] since they are valid for a larger range of variable z (in particular, in domains which contain both the origin and infinity), as well as a larger range of the parameter a, and also include explicit (computable) error bounds. Since the results of this paper, when taken together, are valid for all (unbounded) complex values of z, they can be used to obtain uniform asymptotic approximations of the zeros of $\theta_n(z;a)$ (or $y_n(z;a)$). For each n and a, these zeros lie on a certain curve in the complex plane (which we briefly discuss in section 5), and for earlier results concerning these zeros, see [3], [5], [6], and [11].

Before proceeding, we present some important properties of the reverse Bessel polynomials. Many of these results can be found by perusing the literature, most notably the comprehensive monograph of Grosswald [6].

First, the polynomials are related to the confluent hypergeometric function

$$U(a,c,z) = \frac{1}{\Gamma(a)} \int_0^\infty t^{a-1}(1+t)^{c-a-1}e^{-zt}dt \quad \left(|\arg(z)| < \frac{1}{2}\pi, \quad \operatorname{Re} a > 0 \right) \tag{1.5}$$

by the relation

$$\theta_n(z;a) = 2^{n+a-1}z^{2n+a-1}U(n+a-1, 2n+a, 2z). \tag{1.6}$$

Also, in terms of the Laguerre polynomials $L_n^{(\alpha)}(x)$ they are expressible as

$$\theta_n(z;a) = (-1)^n n! 2^{-n} L_n^{(-2n-a+1)}(2z). \tag{1.7}$$

A Rodrigues formula is given by

$$\theta_n(z;a) = (-1)^n 2^{-n} e^{2z} z^{2n+a-1} \frac{\mathrm{d}^n}{\mathrm{d}z^n}(z^{-n-a+1} e^{-2z}). \tag{1.8}$$

The following provides a generating function expansion:

$$\left\{\frac{1+(1-2t)^{1/2}}{2}\right\}^{2-a} \frac{\exp\{z(1-(1-2t)^{1/2})\}}{(1-2t)^{1/2}} = \sum_{k=0}^{\infty} \frac{t^k \theta_k(z;a)}{k!}. \tag{1.9}$$

A recurrence relation (in the order n) is given by the relation

$$(n+a-1)(2n+a-2)\theta_{n+1}(z;a) \tag{1.10}$$

$$= \left\{(2n+a)\left(n-1+\frac{1}{2}a\right)+(a-2)z\right\}(2n+a-1)\theta_n(z;a)+n(2n+a)z^2\theta_{n-1}(z;a).$$

Special values worth noting are

$$\theta_n(0;a) = \frac{\Gamma(2n+a-1)}{2^n \Gamma(n+a-1)} \tag{1.11}$$

and

$$\theta_n(z;2) \equiv \theta_n(z) = \sqrt{\frac{2}{\pi}} z^{n+1/2} e^z K_{n+1/2}(z). \tag{1.12}$$

In the latter equation, $K_\nu(z)$ denotes the modified Bessel function, and the polynomials $\theta_n(z)$ in this equation are known as Bessel polynomials; indeed, it is from the relation (1.12) that the terminology *generalized Bessel polynomial* arises.

Finally, the reverse generalized Bessel polynomials $\theta_n(z;a)$ satisfy the following second order linear differential equation:

$$z\frac{\mathrm{d}^2\theta}{\mathrm{d}z^2} - (2n-2+a+2z)\frac{\mathrm{d}\theta}{\mathrm{d}z} + 2n\theta = 0. \tag{1.13}$$

It is from this equation, suitably transformed, that we shall derive our asymptotic results.

The plan of this paper is as follows. In section 2 we define two companion functions to $\theta_n(z;a)$, which are also solutions of (1.13). These are characterized as being recessive at $z=\infty$ in the left half plane ($\frac{1}{2}\pi \le \arg(z) \le \frac{3}{2}\pi$), and at $z=0$, respectively. Since $\theta_n(z;a)$ is the solution of (1.13) that is recessive at $z=\infty$ in the right half plane, the three solutions form a numerically satisfactory set for the whole complex z plane (see [8, Chap. 5, sect. 7]). In section 3 of this paper a Liouville transformation is given which transforms the differential equation (1.13) into a canonical form from which Airy function expansions can be derived. These expansions are given in section 4 and are then identified with $\theta_n(z;a)$, and the two other solutions as defined in section 2. The main result for $\theta_n(z;a)$ is given by Theorem 4.1 in section 4. Simpler expansions, involving elementary (exponential) functions are given by applying the

well-known Liouville–Green theory for ordinary differential equations in the complex plane. In section 5 such expansions are given which are valid for complex z (but in more restricted domains than those of section 4). The results of section 5 are then simplified even further in sections 6 and 7, where z is assumed to be real (positive in section 6, and negative in section 7).

When $a = 2$ the results of section 4 are equivalent to the special case of the asymptotic expansion of Bessel functions in the complex plane given by Olver [8, Chap. 11, sect. 10]. The reader may find it helpful to compare the results of Olver to those of section 4 of the present paper.

Throughout this paper z and a take real or complex values (as specified), but we assume that n (which is large) is a positive integer. It is straightforward to extend the following results to noninteger positive values of n (with $\theta_n(z;a)$, no longer a polynomial, defined by (1.6)), but we do not pursue this.

2. Companion solutions of the differential equation. The first step in obtaining asymptotic solutions to (1.13) is to remove the first derivative. To this end, it is straightforward to show that if $\theta(z)$ is any solution of (1.13), then

$$w(z) = z^{1-n-a/2}e^{-z}\theta(z) \tag{2.1}$$

satisfies the differential equation

$$\frac{\mathrm{d}^2w}{\mathrm{d}z^2} = \left\{1 + \frac{a-2}{z} + \frac{(2n+a)(2n+a-2)}{4z^2}\right\} w. \tag{2.2}$$

We shall define three fundamental solutions of the differential equation (2.2), which will be characterized as being recessive at certain points in the complex plane, and indeed will form a numerically satisfactory set for the whole complex plane (subject to appropriate restrictions on $\arg(z)$).

By introducing (for convenience) a normalizing constant 2^{-n-a+1}, we have from (2.1) as the first of these three solutions

$$w_n^{(0)}(z;a) = 2^{-n-a+1}z^{1-n-a/2}e^{-z}\theta_n(z;a), \tag{2.3}$$

or alternatively, from (1.6),

$$w_n^{(0)}(z;a) = e^{-z}z^{n+a/2}U(n+a-1, 2n+a, 2z). \tag{2.4}$$

Now from (1.1), (1.4), and (2.3) we observe that as $z \to \infty$

$$w_n^{(0)}(z;a) = 2^{-n-a+1}z^{1-a/2}e^{-z}\{1 + O(z^{-1})\}. \tag{2.5}$$

The significance of (2.5) is that $w_n^{(0)}(z;a)$ is seen to be recessive in the right half plane. To obtain a solution which is recessive in the left half plane we replace U by V in (2.4), where V denotes the confluent hypergeometric function defined by [8, Chap. 7, (10.03)]. Thus we have as our second fundamental solution

$$w_n^{(1)}(z;a) = (-1)^{n+1}e^{-z}z^{n+a/2}V(n+a-1, 2n+a, 2z), \tag{2.6}$$

where again for convenience we have introduced a constant (this time the factor $(-1)^{n+1}$). A branch of V in the definition given by [8, Chap. 7, (10.03)] must be specified, and we choose it so that $\arg(-z) = \pi$ when $\arg(z) = 0$. Thus we have in effect defined V by the relation

$$V(a,c,z) = e^zU(c-a, c, ze^{-\pi i}). \tag{2.7}$$

Next, from (2.6) and (2.7), we see that

$$w_n^{(1)}(z;a) = (-1)^{n+1} e^z z^{n+a/2} U(n+1, 2n+a, 2ze^{-\pi i}), \tag{2.8}$$

and hence from [8, Chap. 7, (10.01)]

$$w_n^{(1)}(z;a) = 2^{-n-1} z^{a/2-1} e^z \{1 + O(z^{-1})\} \tag{2.9}$$
$$\left(z \to \infty \text{ with } -\frac{1}{2}\pi + \delta \le \arg(z) \le \frac{5}{2}\pi - \delta\right).$$

This confirms that $w_n^{(1)}(z;a)$ is recessive in the left half plane $\frac{1}{2}\pi \le \arg(z) \le \frac{3}{2}\pi$.

Our third of the numerically satisfactory set of solutions is to be recessive at the regular singularity at $z = 0$. We define it by

$$w_n^{(-1)}(z;a) = e^{-z} z^{n+a/2} \mathbf{M}(n+a-1, 2n+a, 2z), \tag{2.10}$$

where $\mathbf{M}$ is the confluent hypergeometric function defined by [8, Chap. 7, (9.04)], that is,

$$\mathbf{M}(a,c,z) = \frac{1}{\Gamma(c)} {}_1F_1(a;c;z) = \sum_{s=0}^{\infty} \frac{(a)_s}{s!\Gamma(c+s)} z^s. \tag{2.11}$$

It is then seen from (2.10) and (2.11) that

$$w_n^{(-1)}(z;a) = \frac{z^{n+a/2}}{\Gamma(2n+a)}\{1 + O(z)\} \quad (z \to 0). \tag{2.12}$$

An important connection formula relating the three solutions can be derived from well-known results concerning confluent hypergeometric functions (see [8, Chap. 7, (10.10)]), and for the case where n is an integer (as is assumed throughout this paper), we arrive at

$$w_n^{(-1)}(z;a) = (-1)^{n+1} \frac{e^{a\pi i}}{n!} w_n^{(0)}(z;a) + \frac{1}{\Gamma(n+a-1)} w_n^{(1)}(z;a). \tag{2.13}$$

We finally remark that all three solutions can be regarded as generalizations of the standard modified Bessel functions, since, on setting $a = 2$ in (2.4), (2.8), and (2.10) and referring to [8, Chap. 7, Exercises 9.1 and 10.1], we find that

$$w_n^{(-1)}(z;2) = \sqrt{\frac{\pi}{2}}\ \frac{z^{1/2}}{2^n \Gamma(n+1)} I_{n+1/2}(z), \tag{2.14}$$

$$w_n^{(0)}(z;2) = \frac{z^{1/2}}{\sqrt{2\pi} 2^n} K_{n+1/2}(z), \tag{2.15}$$

and

$$w_n^{(1)}(z;2) = -i \frac{z^{1/2}}{\sqrt{2\pi} 2^n} K_{n+1/2}(ze^{-\pi i}). \tag{2.16}$$

The relation (2.15), of course, could also have been deduced from (1.12) and (2.3).

3. Preliminary transformations for the turning point case. It is easy to see that the right-hand side of (2.2) has two zeros, which as $n \to \infty$ are located at points given by $z = \pm in + O(1)$. These zeros are the so-called turning points of the equation, and we wish for them to be bounded. With this in mind we shall rescale the independent variable. Before doing so, we define as our large parameter

$$u = n + \frac{1}{2}, \tag{3.1}$$

and, for convenience, we introduce a parameter α by the relation $a = 2 + u\alpha$, i.e.,

$$\alpha = \frac{a-2}{u}. \tag{3.2}$$

We shall assume throughout this paper that α is bounded and not close to -1. Some of the subsequent results also break down when α is close to -2. The reason for these two values being exceptional will become clear shortly. Thus the original parameter a can range from 0 to the same magnitude as u in absolute value, but it must be bounded away from $-u$ (and for the most part $-2u$): more precisely, we require that

$$|a| = O(u), \quad |a+u| \geq \delta u > 0, \text{ and } |a+2u| \geq \delta u > 0. \tag{3.3}$$

Now, on replacing z by uz in (2.2), we are led to the following equation which will be the focus of our attention, and which is satisfied by each function $w_n^{(j)}(uz; a)$ $(j = 0, \pm 1)$:

$$\frac{\mathrm{d}^2 w}{\mathrm{d}z^2} = \{u^2 f(\alpha, z) + g(z)\} w, \tag{3.4}$$

where

$$f(\alpha, z) = 1 + \frac{\alpha}{z} + \frac{(2+\alpha)^2}{4z^2} \tag{3.5}$$

and

$$g(z) = -\frac{1}{4z^2}. \tag{3.6}$$

This particular partitioning was chosen so that the subsequent expansions will be uniformly valid at both $z = 0$ and $z = \infty$. To see why this is so, we refer the reader to [8, Chap. 6, sect.s 4.3 and 5.3].

We are considering the case where $u \to \infty$, and as such the dominant term is $u^2\ f(\alpha, z)$, except near the zeros of $f(\alpha, z)$. The function $f(\alpha, z)$ has two zeros, say, at $z_1(\alpha)$ and $z_2(\alpha)$, where

$$z_{1,2}(\alpha) = \pm i\sqrt{1+\alpha} - \frac{1}{2}\alpha, \tag{3.7}$$

and consequently, $f(\alpha, z)$ can be expressed in the form

$$f(\alpha, z) = \frac{(z - z_1)(z - z_2)}{z^2}. \tag{3.8}$$

The zeros $z_1(\alpha)$ and $z_2(\alpha)$ are the turning points of (3.4). The principal branch of the square root in (3.7) is taken for both, so that when α is real and greater than -1

the turning point $z_1(\alpha)$ lies in the upper half plane (second quadrant when $\alpha > 0$), and the turning point $z_2(\alpha)$ lies in the lower half plane (third quadrant when $\alpha > 0$), with both being continuous functions of α in the complex plane having a cut from $\alpha = -1$ to $\alpha = -\infty$. When α is real and greater than -1 we observe that

$$|z_{1,2}(\alpha)| = 1 + \frac{1}{2}\alpha. \tag{3.9}$$

Note that as $\alpha \to 0$ (for example, when $n \to \infty$ with a fixed),

$$z_{1,2}(\alpha) = \pm i - \frac{1}{2}(1 \mp i)\alpha \mp \frac{1}{8}i\alpha^2 + O(\alpha^3). \tag{3.10}$$

Thus from (3.1), (3.2), (3.3), (3.7), and (3.10) we perceive that these turning points are bounded, and if a is fixed, they approach the points $z = \pm i$, respectively, when $u \to \infty$. Also, in the special case $-1 < \alpha < \infty$ these turning points are complex conjugates, located on the line $\operatorname{Re} z = -\frac{1}{2}\alpha$. At the end of this section we discuss in more detail the position of the two turning points in the complex plane.

Our aim is to obtain asymptotic solutions which are valid in domains which contain the turning point $z = z_1(\alpha)$ (as well as $z = 0$ and $z = \infty$). The domains will be specified precisely later and will be seen to depend on the value of the parameter α. If α lies in the interval $-1 < \alpha < \infty$ the domains of validity will be shown to contain the whole upper half plane (as well as parts of the lower half plane).

To obtain the desired asymptotic expansions we shall use a Liouville transformation (see [8, Chap. 6, (1.03) and (1.04)]) to transform (3.4) into the form

$$\frac{\mathrm{d}^2 W}{\mathrm{d}\zeta^2} = \{u^2\zeta + \psi(\zeta)\}W, \tag{3.11}$$

where W and ζ are, respectively, new dependent and independent variables. The new differential equation (3.11) has a turning point at $\zeta = 0$, which must correspond to the original turning point at $z = z_1$. The nondominant term, the so-called Schwarzian derivative $\psi(\zeta)$ (given by (3.13) and (3.14) below), will then be analytic at $\zeta = 0$. A general asymptotic theory for differential equations of the form (3.11), for the case ζ complex, is then supplied by Olver [8, Chap. 11, Thm. 9.1].

The appropriate Liouville transformation is given by [8, Chap. 11, (3.02)]

$$\zeta\left(\frac{\mathrm{d}\zeta}{\mathrm{d}z}\right)^2 = f(\alpha, z), \qquad W = (f/\zeta)^{1/4}w, \tag{3.12}$$

and this leads to

$$\psi(\zeta) = \frac{5}{16\zeta^2} + \frac{\zeta\{4ff'' - 5f'^2\}}{16f^3} + \frac{\zeta g}{f}, \tag{3.13}$$

or, equivalently (from (3.5) and (3.6)),

$$\psi(\zeta) = \frac{5}{16\zeta^2} - \frac{\zeta z}{16(z - z_1)^3(z - z_2)^3}[4z^3 - (4 + 3\alpha)(4 + \alpha)z - \alpha(2 + \alpha)^2]. \tag{3.14}$$

Integration of the first equation of (3.12) yields the relationship

$$\frac{2}{3}\zeta^{3/2} = \int_{z_1}^{z} f^{1/2}(\alpha, t)\,\mathrm{d}t = \int_{z_1}^{z} \frac{(t - z_1)^{1/2}(t - z_2)^{1/2}}{t}\mathrm{d}t, \tag{3.15}$$

where the lower integration limit is chosen to ensure that $\zeta = 0$ corresponds to $z = z_1(\alpha)$. As a result ζ, regarded as a function of z, is analytic at $z = z_1$ but has branch points at $z = 0$ and $z = z_2$. Explicit integration of the right-hand side of (3.15) yields

$$\begin{aligned}\frac{2}{3}\zeta^{3/2} &= [(z-z_1)(z-z_2)]^{1/2} \\ &\quad - \frac{1}{2}(z_1+z_2)\ln\left\{\frac{z-\frac{1}{2}(z_1+z_2)+[(z-z_1)(z-z_2)]^{1/2}}{(z_1-z_2)/(2i)}\right\} \\ &\quad + \sqrt{z_1z_2}\ln\left\{\frac{[z_1z_2(z-z_1)(z-z_2)]^{1/2}+\frac{1}{2}z(z_1+z_2)-z_1z_2}{z(z_1-z_2)/(2i)}\right\} - \frac{1}{2}(1+\alpha)\pi i.\end{aligned} \tag{3.16}$$

In this relationship it is helpful to note from (3.7) that

$$z_1z_2 = \left(1+\frac{1}{2}\alpha\right)^2, \quad z_1+z_2 = -\alpha, \quad \frac{z_1-z_2}{2i} = \sqrt{1+\alpha}. \tag{3.17}$$

The $z-\zeta$ transformation (3.16) is quite complicated. In order to understand it more fully, and to specify the branches precisely, it is convenient to introduce an intermediate variable ω by

$$\frac{2}{3}\zeta^{3/2} = \omega = \int_{z_1}^{z}\frac{(t-z_1)^{1/2}(t-z_2)^{1/2}}{t}\mathrm{d}t. \tag{3.18}$$

We then introduce a branch cut along the ray emanating from $z = 0$ and passing through $z = z_2$. In addition, since ω (regarded as a function of z) has a branch point at $z = z_1$, we introduce (temporarily) a branch cut from $z = z_1$ to $z = \infty$ along the ray $\arg(z) = \arg(z_1)$; see Figure 1. Then ω is defined as being a continuous function of z in this cut plane, with branches specified so that $\operatorname{Im}\omega = -\frac{1}{2}(1+\alpha)\pi$ when $\alpha \in (-1,\infty)$ and z is real. Assuming for the moment that α is real and lying in the interval $-1 < \alpha < \infty$, we find that (on utilizing (3.16))

$$\omega = z + \frac{1}{2}\alpha\ln(2z) + \frac{1}{2}\alpha - \frac{1}{2}(1+\alpha)\ln(1+\alpha) - \frac{1}{2}(1+\alpha)\pi i + O(z^{-1}) \tag{3.19}$$

as $z \to \infty$ in the right half plane, where $\ln(2z)$ is real when z is positive. For each fixed z in the cut plane, we also specify that ω depends continuously on α in the complex plane having a cut from $\alpha = -1$ to $\alpha = -\infty$.

From these definitions we observe that as $\operatorname{Re} z \to -\infty$ (to the left of the rays ABC and DEF in Figure 1)

$$\omega = -z - \frac{1}{2}\alpha\ln(2z) + \frac{1}{2}(1+\alpha)\ln(1+\alpha) - \frac{1}{2}\alpha + \frac{1}{2}(1+\alpha)\pi i + O(z^{-1}). \tag{3.20}$$

With our choice of branches the effect of the $z-\omega$ transformation is depicted in Figures 1, 2(a), and 2(b) (for the case $\alpha > 0$). Figure 2(a) depicts the region mapped to the right of the rays $\arg z = \arg z_1$ and $\arg z = \arg z_2$, and Figure 2(b) depicts the region mapped to the left of these rays. In these (and subsequent figures) corresponding points are labeled by the same capital letters. For example, the point $\omega = -(1+\alpha)\pi i$ corresponds to $z = z_2$ on the right-hand side of the cut, and is labeled

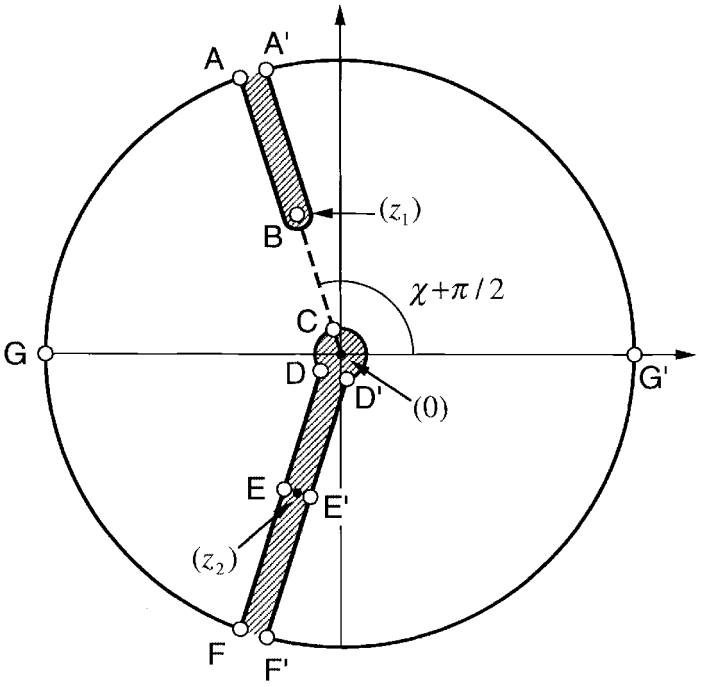


FIG. 1. *z plane.*

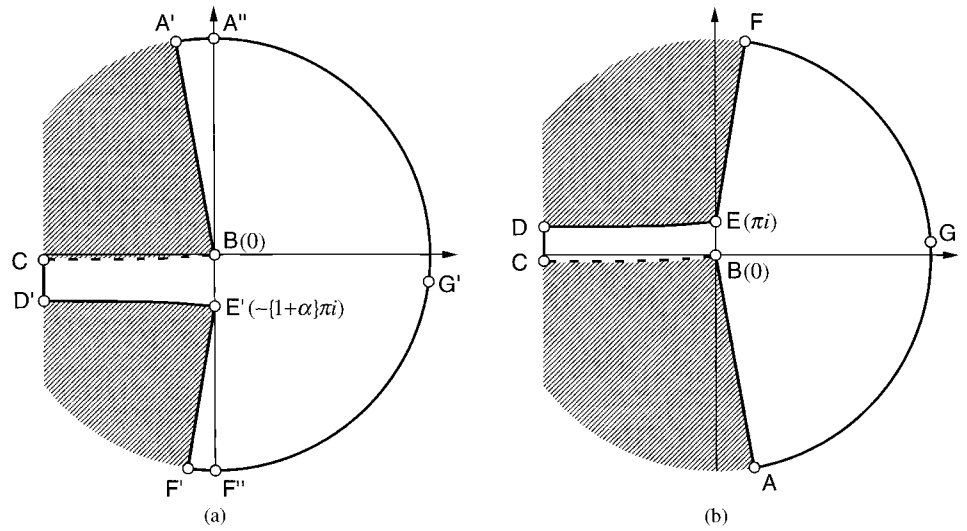


FIG. 2. (a) *ω plane,* (b) *ω plane.*

E′ in Figures 1 and 2(a). Likewise, $z = z_2$ on the other side of the cut is mapped to $\omega = \pi i$ (labeled E in Figures 1 and 2(b)).

Finally, from the relationship $\omega = \frac{2}{3}\zeta^{3/2}$, the domain, say, Δ, corresponding to the cut z plane of Figure 3 (the cut emanating from $z = z_1$ now being removed) can be determined via Figures 2(a) and 2(b). This domain is shown in Figure 4, with

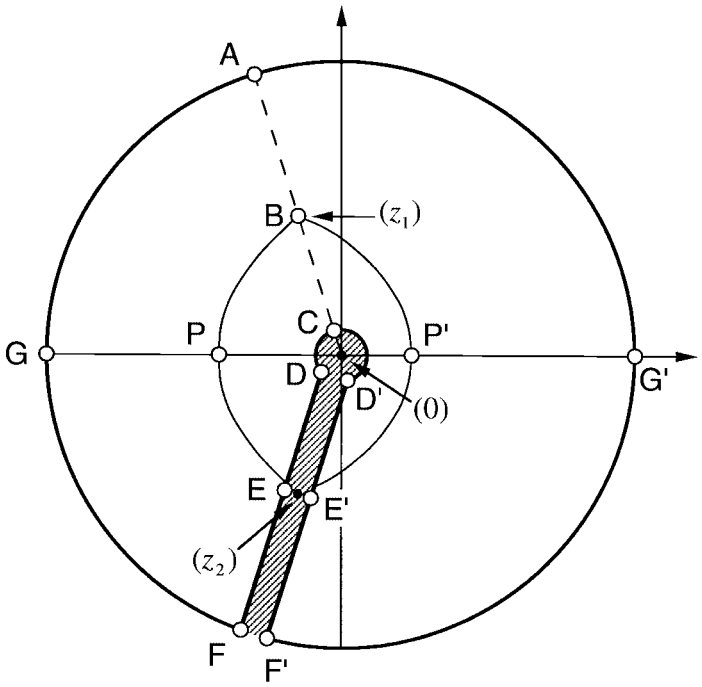


FIG. 3. *z plane.*

the points E and E′, which correspond to the turning point $z = z_2$ on either side of the cut, being mapped to the points $\zeta = e^{-\pi i/3}[\frac{3}{2}(1+\alpha)\pi]^{2/3}$ and $\zeta = -(\frac{3}{2}\pi)^{2/3}$, respectively. The origin in the ζ plane, labeled B, corresponds to the turning point $z = z_1$, and is, of course, a regular point of the $z - \zeta$ transformation.

We shall require the following asymptotic results for the cases $z \to \infty$ and $z \to 0$. First, exponentiation of (3.19) yields

$$\exp\left\{\frac{2}{3}u\zeta^{3/2}\right\} = e^{-u(1+\alpha)\pi i/2}(1+\alpha)^{-u(1+\alpha)/2}e^{u\alpha/2}e^{uz}(2z)^{u\alpha/2}\{1+O(z^{-1})\} \tag{3.21}$$

as $z \to \infty$: here and in section 4, this means that $z \to \infty$ in *any* direction in the cut plane depicted in Figure 3. However, when $z \to \infty$ it is understood that ζ approaches infinity in the part of Δ, containing the positive real axis, which is bounded by the curves BEF and BE′F′ (see Figure 4). In (3.21) $\exp\{\frac{2}{3}u\zeta^{3/2}\}$ takes its principle value in the right half plane is continuous elsewhere in the stated subdomain.

Next, we can show from (3.16) that

$$\begin{aligned}\frac{2}{3}\zeta^{3/2} &= \frac{1}{2}(2+\alpha)\ln\{2(2+\alpha)^{-2}z\} + 1 + \frac{1}{2}\alpha \\ &\quad + \frac{1}{2}(1+\alpha)\ln(1+\alpha) - \frac{1}{2}(1+\alpha)\pi i + O(z)\end{aligned} \tag{3.22}$$

as $z \to 0$. Note that $\zeta \to \infty e^{-2\pi i/3}$ in this case. In both (3.21) and (3.22) the O terms are real when z is positive and α is lying in the interval $-1 < \alpha < \infty$.

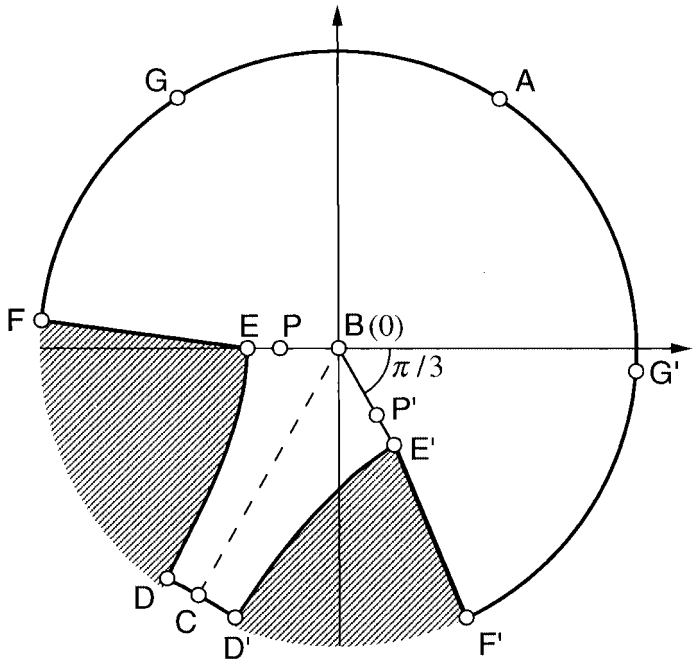


FIG. 4. ζ *plane.*

It is also worth noting that when $-1 < \alpha < \infty$

$$\zeta = \frac{2e^{-\pi i/6-2i\chi(\alpha)/3}(1+\alpha)^{1/6}}{(2+\alpha)^{2/3}}(z-z_1) + O\{(z-z_1)^2\}, \tag{3.23}$$

as $z \to z_1$, where (see Figure 1)

$$\chi(\alpha) = \arg(z_1(\alpha)) - \frac{1}{2}\pi = \sin^{-1}\left(\frac{\alpha}{2+\alpha}\right). \tag{3.24}$$

The principal value of the inverse sine applies here, so that as $\alpha \to 0$

$$\chi(\alpha) = \frac{1}{2}\alpha + O(\alpha^2). \tag{3.25}$$

We conclude this section by examining the location of the two turning points in the z plane when α is complex. First, let us write

$$\alpha = -1 + r^2 e^{i\theta}, \tag{3.26}$$

where $r > 0$ and $-\pi \le \theta \le \pi$. Then, from (3.7) and (3.26) one finds that

$$\text{Im } z_1(\alpha) = r\cos\left(\frac{1}{2}\theta\right)\left\{1 - r\sin\left(\frac{1}{2}\theta\right)\right\} \tag{3.27}$$

and

$$\text{Im } z_2(\alpha) = -r\cos\left(\frac{1}{2}\theta\right)\left\{1 + r\sin\left(\frac{1}{2}\theta\right)\right\}. \tag{3.28}$$

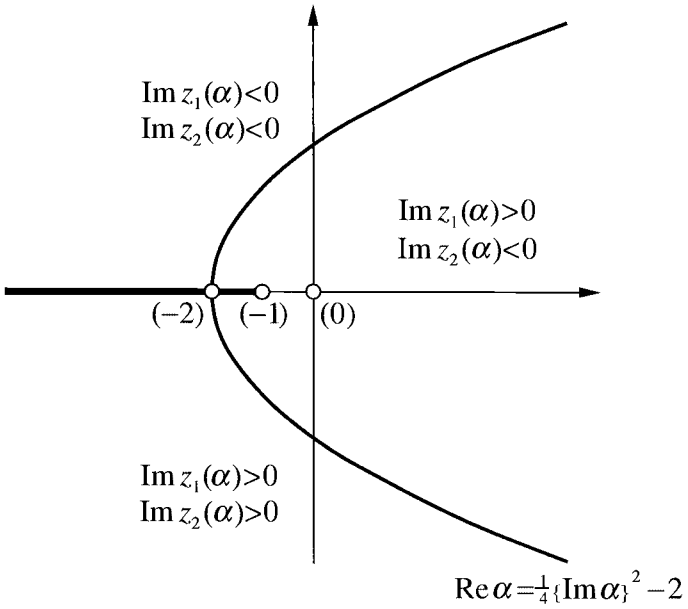


FIG. 5. α *plane.*

Hence, Im $z_1(\alpha) > 0$ when $-\pi < \theta \leq 0$ (α in lower half plane), and Im $z_2(\alpha) < 0$ when $0 \leq \theta < \pi$ (α in upper half plane). On the other hand $z_1(\alpha)$ passes from the upper to lower half planes as α passes across the curve $r\sin(\frac{1}{2}\theta) = 1$, $0 < \theta < \pi$; from (3.26) we see that this is the curve Im $\sqrt{\alpha+1} = 1$, i.e., the half-parabola Re $\alpha = \frac{1}{4}\{\text{Im}\,\alpha\}^2 - 2$, Im $\alpha > 0$. On this half-parabola $z_1(\alpha)$ is real and negative, taking the value $z_1 = -\frac{1}{8}\{\text{Im}\,\alpha\}^2$. Likewise, $z_2(\alpha)$ passes from the lower to upper half planes as α passes across the half-parabola Re $\alpha = \frac{1}{4}\{\text{Im}\,\alpha\}^2 - 2$, Im $\alpha < 0$ (taking the value $z_2 = -\frac{1}{8}\{\text{Im}\,\alpha\}^2$ on this curve). All these situations are depicted in Figure 5.

Consider now the case when α lies on the interval $-\infty < \alpha \leq -1$, say, below the cut. We set $e^{i\theta} = -1$ in (3.26), so that $\alpha = -1 - r^2$ ($0 \leq r < \infty$). Then, it follows that $z_1(\alpha) = \frac{1}{2}(1+r)^2$ and $z_2(\alpha) = \frac{1}{2}(1-r)^2$ (or vice versa if α lies above the cut). Thus, both turning points lie on the nonnegative real z axis, with $r = 0$ ($\alpha = -1$) and $r = 1$ ($\alpha = -2$) being critical values: in the former case the two turning points coalesce with one another at the point $z = \frac{1}{2}$, and in the latter case the turning point $z_2(\alpha)$ coalesces with the pole $z = 0$ (or alternatively, $z_1(\alpha)$ coalesces with the pole when $\alpha \to -2$ above the cut). Both the cases $\alpha = -1$ and $\alpha = -2$ are beyond the scope of this paper, although they can be tackled using established results. A general theory for two coalescing turning points (in the real variable case) is provided by [9], and a general theory for a coalescing turning point and double pole (in both real and complex variable cases) is provided by [2]. We remark, however, that some of the Airy function results that follow in section 4 will still be valid when $\alpha \to -2$, since the turning point $z_1(\alpha)$ (defined as $z_1(\alpha) = \frac{1}{2}(1+r)^2$ when $\alpha = -1-r^2$) will be bounded away from both the pole at $z = 0$ and the other turning point $z_2(\alpha)$. However, when $\alpha \to -2$ none of the subsequent expansions will be valid in the vicinity of the pole $z = 0$.

4. Uniform asymptotic expansions involving Airy functions. We now focus on the transformed differential equation (3.11). We apply Theorem 9.1 of [8, Chap. 11] (with n replaced by N) to this equation to obtain the following three solutions ($j = 0, \pm 1$):

$$W_{2N+1,j}(u,\zeta) = \mathrm{Ai}_j(u^{2/3}\zeta)\sum_{s=0}^{N}\frac{A_s(\zeta)}{u^{2s}} + \frac{1}{u^2}\frac{\mathrm{d}\ \mathrm{Ai}_j(u^{2/3}\zeta)}{\mathrm{d}\zeta}\sum_{s=0}^{N-1}\frac{B_s(\zeta)}{u^{2s}} + \varepsilon_{2N+1,j}(u,\zeta), \tag{4.1}$$

where $\mathrm{Ai}(z)$ is the standard Airy function and

$$\mathrm{Ai}_j(u^{2/3}\zeta) = \mathrm{Ai}(u^{2/3}\zeta e^{-2\pi i j/3}). \tag{4.2}$$

The significance of each Airy function $\mathrm{Ai}_j(u^{2/3}\zeta)$ ($j = 0, \pm 1$) is that it is recessive (exponentially small) inside the sector $|\arg(\zeta e^{-2\pi j i/3})| \leq \frac{1}{3}\pi$, and dominant (exponentially large) outside this sector (except on the boundaries of the three sectors). Following [8, Chap. 11, sect. 8.1] we denote these three sectors by $\mathbf{S}_j$ ($j = 0, \pm 1$).

The coefficients in the asymptotic expansions (4.1) are defined recursively by $A_0(\zeta) = 1$,

$$A_{s+1}(\zeta) = -\frac{1}{2}B_s'(\zeta) + \frac{1}{2}\int_{\infty}^{\zeta}\psi(\nu)B_s(\nu)\mathrm{d}\nu \quad (s = 0, 1, 2, \ldots) \tag{4.3}$$

and

$$B_s(\zeta) = \frac{1}{2\zeta^{1/2}}\int_0^{\zeta}\{\psi(\nu)A_s(\nu) - A_s''(\nu)\}\frac{\mathrm{d}\nu}{\nu^{1/2}} \quad (s = 0, 1, 2, \ldots). \tag{4.4}$$

The lower integration limits in (4.3) are taken at a point of infinity in the part of Δ which is bounded by the curves BEF and $\mathsf{BE'F'}$ (see Figure 4), i.e., corresponding to $z = \infty$. As a result, from (3.14) and (3.19) and induction on (4.3) and (4.4), it can be shown that for $s = 0, 1, 2 \ldots$

$$\lim_{\substack{\zeta\to\infty\\(z\to\infty)}} A_{s+1}(\zeta) = 0 \tag{4.5}$$

and

$$\lim_{\substack{\zeta\to\infty\\(z\to\infty)}} \zeta^{1/2}B_s(\zeta) = l_s \tag{4.6}$$

for some constants $\{l_s\}_{s=0}^{\infty}$ (which will be discussed later).

The error terms $\varepsilon_{2N+1,j}(u,\zeta)$ ($j = 0, \pm 1$) are $O(u^{-2N-1})$ as $u \to \infty$, uniformly for ζ lying in certain (unbounded) subdomains of Δ, which we denote by $\mathbf{Z}_j(\alpha)$, since these regions of asymptotic validity will depend on α. The domains $\mathbf{Z}_j$ are described in a general setting by Olver in [8, Chap. 11, sect. 9] and are denoted by $\mathbf{Z}_j(u,\alpha_j)$ in that reference: they consist of all points in Δ which can be linked by a certain path $\mathcal{L}_j$ to a so-called reference point in $\Delta \cap \mathbf{S}_j$ (which in [8, Chap. 11, sect. 9] is denoted by $\zeta = \alpha_j$). For our purposes we choose the reference points to be at $\zeta = +\infty$, $\infty e^{2\pi i/3}$, $\infty e^{-2\pi i/3}$ for $j = 0, 1, -1$, respectively. We are permitted to take these points at

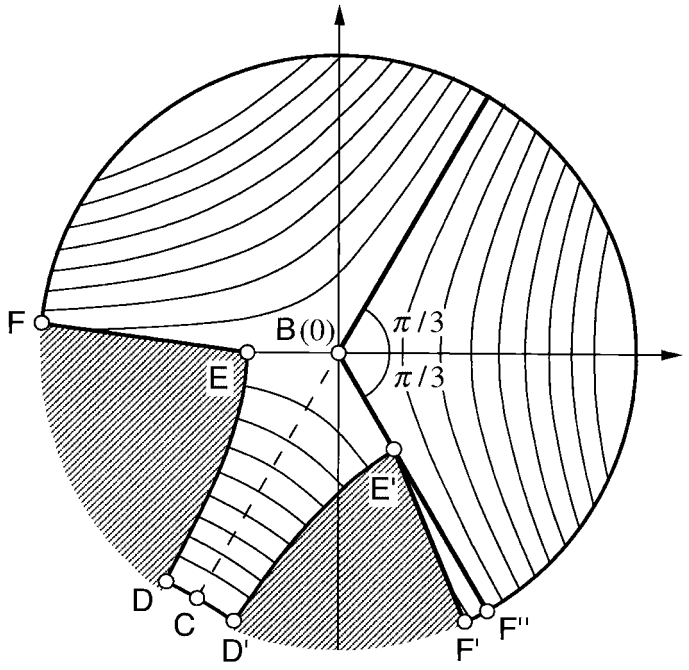


FIG. 6. ζ *plane.*

infinity by virtue of the fact that $\psi^{(s)}(\zeta) = O(\zeta^{-2-s})$ as $\zeta \to \infty$ (for both the cases $z \to 0$ and $z \to \infty$): see (3.14), (3.21), and (3.22), and [8, sect. 9.3].

Each path $\mathcal{L}_j$ must consist of a finite chain of R_2 arcs which lie inside Δ, avoid neighborhoods of the finite singularities $\zeta = e^{-\pi i/3}[\frac{3}{2}(1+\alpha)\pi]^{2/3}$ and $\zeta = -(\frac{3}{2}\pi)^{2/3}$ (labeled E and E' in the figures), and be such that Re $\nu^{3/2}$ is monotonic as ν passes along $\mathcal{L}_j$ from $\infty e^{2\pi i j/3}$ to ζ. In this definition the branch of Re $\nu^{3/2}$, for each $j = 0, \pm 1$, is defined to be continuous with Re $\nu^{3/2} \geq 0$ in $\mathbf{S}_j$ and Re $\nu^{3/2} \leq 0$ in $\mathbf{S}_{j-1} \cup \mathbf{S}_{j+1}$. (Here and elsewhere the suffixes are enumerated modulo 3.)

Although the above descriptions of the domains of validity $\mathbf{Z}_j(\alpha)$ appear quite complicated, the idea is quite straightforward. As an example, suppose that α is real and positive, so that the turning points are located as in Figure 4, with the boundary of the domain Δ also being depicted in this figure. Now, let us first determine the domain $\mathbf{Z}_0(\alpha)$. In Figure 6 the so-called level curves Re $\zeta^{3/2}$ = constant are superimposed on the domain Δ, with Re $\zeta^{3/2} \geq 0$ in $\mathbf{S}_0$ and Re $\zeta^{3/2} \leq 0$ in $\mathbf{S}_1 \cup \mathbf{S}_{-1}$: thus the value of the associated constant for each of these curves is zero on the boundaries of $\mathbf{S}_j$, increases to infinity as $\zeta \to +\infty$, and decreases to minus infinity as $\zeta \to \infty e^{\pm 2\pi i/3}$. It is then readily seen from this figure that all points in Δ (excluding neighborhoods of the singularities at E and E') lie in $\mathbf{Z}_0(\alpha)$, since each point ζ can be linked to the reference point $\zeta = +\infty$ by a path $\mathcal{L}_0$ on which Re $\nu^{3/2}$ is nonincreasing (as ν passes along $\mathcal{L}_0$ from $+\infty$ to ζ).

On the other hand the domains $\mathbf{Z}_1(\alpha)$ and $\mathbf{Z}_{-1}(\alpha)$ differ from $\mathbf{Z}_0(\alpha)$ in the following way. They consist of all points in Δ with neighborhoods of E and E' excluded, but all points on or near the ray $\mathsf{E}'\mathsf{F}''$ must also be excluded. The reason is that any path $\mathcal{L}_1$ (or $\mathcal{L}_{-1}$) linking a point on the ray $\mathsf{E}'\mathsf{F}''$ to $\zeta = \infty e^{2\pi i/3}$ (or $\zeta = \infty e^{-2\pi i/3}$) must necessarily pass through (or close to) the singularity at E', due to the requirement

that Re $\nu^{3/2}$ be monotonic along the path. (Recall that the branch of Re $\nu^{3/2}$, as described above, is chosen differently for the cases $j=1$ and $j=-1$.)

Having described the regions of validity, we apply Olver's theorem to obtain the following error bounds, these being uniformly valid for $\zeta \in \mathbf{Z}_j(\alpha)$:

$$|\varepsilon_{2N+1,j}(u,\zeta)| \le \frac{4v_2 M_{j\pm1}(u^{2/3}\zeta)}{E_j(u^{2/3}\zeta)} \times \exp\left\{\frac{4v_1}{u}\mathcal{V}_{\mathcal{L}_j}(\zeta^{1/2}B_0)\right\}\frac{\mathcal{V}_{\mathcal{L}_j}(\zeta^{1/2}B_n)}{u^{2N+1}} \quad (j=0,\pm1). \tag{4.7}$$

In these bounds

$$E_0(z) = \left|\exp\left(\frac{2}{3}z^{3/2}\right)\right|, \tag{4.8}$$

$$E_j(z) = E_0(ze^{-2\pi ij/3}) \quad (j=\pm1), \tag{4.9}$$

$$M_j(z) = \{E_{j+1}^2(z)|\mathrm{Ai}_{j+1}^2(z)| + E_{j-1}^2(z)|\mathrm{Ai}_{j-1}^2(z)|\}^{1/2}, \tag{4.10}$$

$$v_1 = \sup_{\text{all } z}\{\pi|z|^{1/2}M_j^2(z)\}, \tag{4.11}$$

and

$$v_2 = \sup_{\text{all } z}\{\pi E_{j-1}(z)M_j(z)|z^{1/2}\mathrm{Ai}_{j-1}(z)|\}. \tag{4.12}$$

The branches in (4.8) are chosen so that $E_0(u\zeta^{3/2}) \ge 1$ in $\mathbf{S}_0$ and $E_0(u\zeta^{3/2}) \le 1$ in $\mathbf{S}_1 \cup \mathbf{S}_{-1}$. Hence for $j=\pm1$, from the definition (4.9), $E_j(u\zeta^{3/2}) \ge 1$ in $\mathbf{S}_j$ and $E_j(u\zeta^{3/2}) \le 1$ in $\mathbf{S}_{j-1} \cup \mathbf{S}_{j+1}$. Since $\psi(\zeta) = O(\zeta^{-2})$ as $\zeta \to \infty$ (for both the cases $z \to 0$ and $z \to \infty$) it follows from the bound (4.7), and from section 9.3 of [8], that $\varepsilon_{2N+1,j}(u,\zeta) = E_j^{-1}(u^{2/3}\zeta)M_{j\pm1}(u^{2/3}\zeta)O(\zeta^{-3/2})$ as $\zeta \to \infty e^{2\pi ij/3}$ $(j=0,1,-1)$.

Having derived asymptotic solutions, we match them with the standard solutions of (3.4). On account of uniqueness of recessive solutions, we assert that for $j=0,1,-1$

$$\Lambda_{2N+1,j}w_n^{(j)}(uz;a) = \left\{\frac{z^2\zeta}{(z-z_1)(z-z_2)}\right\}^{1/4} W_{2N+1,j}(u,\zeta), \tag{4.13}$$

where $\Lambda_{2N+1,j}$ are constants.

Let us first determine $\Lambda_{2N+1,0}$. To do this we shall use the well-known behavior of Airy functions of large argument (see, for example, [8, Chap. 11, (1.07)])

$$\mathrm{Ai}(w) = \frac{\exp\{-\frac{2}{3}w^{3/2}\}}{2\sqrt{\pi}w^{1/4}}\{1+O(w^{-1})\}, \tag{4.14}$$

$$\mathrm{Ai}'(w) = -\frac{w^{1/4}\exp\{-\frac{2}{3}w^{3/2}\}}{2\sqrt{\pi}}\{1+O(w^{-1})\} \tag{4.15}$$

as $w \to \infty$ with $|\arg(w)| \le \pi - \delta$. Now using these, while letting $z \to \infty$ in (4.1), and referring to (4.5) and (4.6), we find that

$$\left\{\frac{z^2\zeta}{(z-z_1)(z-z_2)}\right\}^{1/4} W_{2N+1,0}(u,\zeta) = \frac{\exp\{-\frac{2}{3}u\zeta^{3/2}\}}{2\sqrt{\pi}u^{1/6}}\left[1-\sum_{s=0}^{N-1}\frac{l_s}{u^{2s+1}}\right]\{1+O(\zeta^{-3/2})\}. \tag{4.16}$$

On the other hand, from (2.5)

$$w_n^{(0)}(uz;a) = 2^{-u-u\alpha-1/2}u^{-u\alpha/2}z^{-u\alpha/2}e^{-uz}\{1+O(z^{-1})\}, \tag{4.17}$$

as $z \to \infty$. Therefore, from (4.13), (4.16), and (4.17) we find that

$$\Lambda_{2N+1,0} = \frac{2^{u+u\alpha-1/2}u^{u\alpha/2-1/6}}{\sqrt{\pi}}\left[1-\sum_{s=0}^{N-1}\frac{l_s}{u^{2s+1}}\right]\lim_{z\to\infty} z^{u\alpha/2}\exp\left\{uz-\frac{2}{3}u\zeta^{3/2}\right\}, \tag{4.18}$$

which, on referring to (3.21), yields our desired formula

$$\Lambda_{2N+1,0} = \frac{e^{u(1+\alpha)\pi i/2}2^{u+u\alpha/2-1/2}u^{u\alpha/2-1/6}(1+\alpha)^{u(1+\alpha)/2}e^{-u\alpha/2}}{\sqrt{\pi}}\left[1-\sum_{s=0}^{N-1}\frac{l_s}{u^{2s+1}}\right]. \tag{4.19}$$

We find $\Lambda_{2N+1,1}$ similarly. When $j=1$ the left-hand side of (4.13) has the form

$$\begin{aligned}&\left\{\frac{z^2\zeta}{(z-z_1)(z-z_2)}\right\}^{1/4} W_{2N+1,1}(u,\zeta)\\ &\quad= \frac{e^{\pi i/6}\exp\{\frac{2}{3}u\zeta^{3/2}\}}{2\sqrt{\pi}u^{1/6}}\left[1+\sum_{s=0}^{N-1}\frac{l_s}{u^{2s+1}}\right]\{1+O(\zeta^{-3/2})\}\end{aligned} \tag{4.20}$$

when $z \to \infty$. Next, from (2.9) we see that

$$w_n^{(1)}(uz;a) = 2^{-u-1/2}u^{u\alpha/2}z^{u\alpha/2}e^{uz}\{1+O(z^{-1})\} \tag{4.21}$$

as $z \to \infty$ (with $-\frac{1}{2}\pi+\delta \le \arg(z) \le \frac{5}{2}\pi-\delta$), and hence, using these in (4.13) (with $j=1$), and again referring to (3.21), we find that

$$\Lambda_{2N+1,1} = \frac{e^{\pi i/6}e^{-(1+\alpha)\pi i/2}2^{u+u\alpha/2-1/2}e^{u\alpha/2}}{\sqrt{\pi}u^{u\alpha/2+1/6}(1+\alpha)^{u(1+\alpha)/2}}\left[1+\sum_{s=0}^{N-1}\frac{l_s}{u^{2s+1}}\right]. \tag{4.22}$$

We finally find $\Lambda_{2N+1,-1}$ by setting $j=-1$ in (4.13) and again letting $z \to \infty$. Thus, we find that

$$\begin{aligned}&\left\{\frac{z^2\zeta}{(z-z_1)(z-z_2)}\right\}^{1/4} W_{2N+1,-1}(u,\zeta)\\ &\quad= \frac{e^{-\pi i/6}\exp\{\frac{2}{3}u\zeta^{3/2}\}}{2\sqrt{\pi}u^{1/6}}\left[1+\sum_{s=0}^{N-1}\frac{l_s}{u^{2s+1}}+\delta_{2N+1}\right]\{1+O(\zeta^{-3/2})\},\end{aligned} \tag{4.23}$$

where

$$\delta_{2N+1} = e^{\pi i/6}2\sqrt{\pi}u^{1/6}\lim_{\zeta\to+\infty}\exp\left\{-\frac{2}{3}u\zeta^{3/2}\right\}\zeta^{1/4}\varepsilon_{2N+1,-1}(u,\zeta). \tag{4.24}$$

Explicit bounds for this constant are available via (4.7), and we note that $\delta_{2N+1} = O(u^{-2N-1})$ as $u \to \infty$. On the other hand, on utilizing (2.5), (2.9), (2.13), (3.1),

(3.2), and (3.21), we find that as $z \to \infty$ with $-\frac{1}{2}\pi+\delta \le \arg(z) \le \frac{3}{2}\pi-\delta$ (i.e., $\zeta \to \infty$ with $|\arg(\zeta e^{-\pi i/3})| \le \frac{2}{3}\pi-\delta$)

$$w_n^{(-1)}(uz;a) = i\,\frac{e^{u(1+\alpha)\pi i/2}e^{u\alpha/2}\exp\{-\frac{2}{3}u\zeta^{3/2}\}}{2^{u+u\alpha/2+1/2}u^{u\alpha/2}(1+\alpha)^{u(1+\alpha)/2}\Gamma(u+\frac{1}{2})}\{1+O(\zeta^{-3/2})\} \tag{4.25}$$
$$+\,\frac{e^{u(1+\alpha)\pi/2}u^{u\alpha/2}(1+\alpha)^{u(1+\alpha)/2}\exp\{\frac{2}{3}u\zeta^{3/2}\}}{2^{u+u\alpha/2+1/2}e^{u\alpha/2}\Gamma(u+u\alpha+\frac{1}{2})}\{1+O(\zeta^{-3/2})\}.$$

In particular, if $\zeta \to +\infty$ the first term on the right-hand side is negligible, and hence from (4.13) and (4.23) we obtain the desired expression

$$\Lambda_{2N+1,-1} = \frac{e^{-\pi i/6}e^{-u(1+\alpha)\pi i/2}2^{u+u\alpha/2-1/2}\Gamma(u+u\alpha+\frac{1}{2})e^{u\alpha/2}}{\sqrt{\pi}u^{1/6}u^{u\alpha/2}(1+\alpha)^{u(1+\alpha)/2}} \tag{4.26}$$
$$\times\left[1+\sum_{s=0}^{N-1}\frac{l_s}{u^{2s+1}}+\delta_{2N+1}\right].$$

Although it is possible to find $\Lambda_{2N+1,-1}$ without the unknown constant δ_{2N+1} by instead letting $\zeta \to \infty e^{-2\pi i/3}$ (i.e., $z \to 0$), we prefer to again let $\zeta \to +\infty$ (i.e., $z \to \infty$) so as to obtain an expression for $\Lambda_{2N+1,-1}$ which involves the same constants $\{l_s\}_{s=0}^{\infty}$ as in (4.19) and (4.22) above. Moreover, from (4.26) we are able to obtain as a byproduct a method for easily calculating these constants, as we now demonstrate. First, from (4.1), (4.2), (4.5), (4.6), and [8, Chap. 11, (1.08) and (1.09)] it can be seen that

$$\left\{\frac{z^2\zeta}{(z-z_1)(z-z_2)}\right\}^{1/4} W_{2N+1,-1}(u,\zeta) \tag{4.27}$$
$$= \frac{e^{\pi i/12}}{\sqrt{\pi}u^{1/6}}\left[\cos\left(i\frac{2}{3}u\zeta^{3/2}+\frac{1}{4}\pi\right)-i\sin\left(i\frac{2}{3}u\zeta^{3/2}+\frac{1}{4}\pi\right)\sum_{s=0}^{N-1}\frac{l_s}{u^{2s+1}}\right]$$
$$+\left\{\frac{z^2\zeta}{(z-z_1)(z-z_2)}\right\}^{1/4}\varepsilon_{2N+1,-1}(u,\zeta)+O(\zeta^{-3/2})$$

as $\zeta e^{2\pi i/3} \to -\infty$. (Note that (4.14) and (4.15) cannot be used here.) Now in (4.27) let $\zeta = \zeta_p$, where $\frac{2}{3}u\zeta_p^{3/2} = 2p\pi i + \frac{1}{4}\pi i$, where p is a positive integer (and the branch is such that $\arg(\zeta) = \frac{1}{3}\pi$). Then $\cos(i\frac{2}{3}u\zeta_p^{3/2}+\frac{1}{4}\pi) = 1$, $\sin(i\frac{2}{3}u\zeta_p^{3/2}+\frac{1}{4}\pi) = 0$, and $\zeta_p \to \infty e^{\pi i/3}$ as $p \to \infty$. Hence, we see that

$$\lim_{\substack{\zeta=\zeta_p\\ p\to\infty}}\left\{\frac{z^2\zeta}{(z-z_1)(z-z_2)}\right\}^{1/4} W_{2N+1,-1}(u,\zeta) = \frac{e^{\pi i/12}}{\sqrt{\pi}u^{1/6}}\{1+O(u^{-2N-1})\}, \tag{4.28}$$

in which

$$O(u^{-2N-1}) = e^{-\pi i/12}\sqrt{\pi}u^{1/6}\lim_{p\to\infty}\zeta_p^{1/4}\varepsilon_{2N+1,-1}(u,\zeta_p). \tag{4.29}$$

Likewise, from (4.25)

$$\lim_{\substack{\zeta=\zeta_p\\ p\to\infty}} w_n^{(-1)}(uz;a) = \frac{e^{\pi i/4}e^{u(1+\alpha)\pi i/2}e^{u\alpha/2}}{2^{u+u\alpha/2+1/2}u^{u\alpha/2}(1+\alpha)^{u(1+\alpha)/2}\Gamma(u+\frac{1}{2})} \tag{4.30}$$

$$+\frac{e^{\pi i/4}e^{u(1+\alpha)\pi i/2}u^{u\alpha/2}e^{-u\alpha/2}(1+\alpha)^{u(1+\alpha)/2}}{2^{n+u\alpha/2+1}\Gamma(u+u\alpha+\frac{1}{2})}.$$

Thus, inserting (4.28) and (4.30) into (4.13) we find that

$$\Lambda_{2N+1,-1}=\frac{e^{-\pi i/6}e^{-u(1+\alpha)\pi i/2}2^{u+u\alpha/2+1/2}}{\sqrt{\pi}u^{1/6}}\{1+O(u^{-2N-1})\} \tag{4.31}$$
$$\times\left[\frac{e^{u\alpha/2}}{u^{u\alpha/2}(1+\alpha)^{u(1+\alpha)/2}\Gamma(u+\frac{1}{2})}+\frac{u^{u\alpha/2}(1+\alpha)^{u(1+\alpha)/2}}{e^{u\alpha/2}\Gamma(u+u\alpha+\frac{1}{2})}\right]^{-1}.$$

If we now compare (4.26) and (4.31) and note that N is arbitrary, we deduce that (at least formally)

$$1+\sum_{s=0}^{\infty}\frac{l_s}{u^{2s+1}}=2\left[1+\frac{e^{u\alpha}\Gamma(u+u\alpha+\frac{1}{2})}{u^{u\alpha}(1+\alpha)^{u(1+\alpha)}\Gamma(u+\frac{1}{2})}\right]^{-1}. \tag{4.32}$$

We first notice that when $\alpha=0$ the right-hand side reduces identically to 1, and hence, it follows that $l_s=0$ for all s. This is in agreement with the well-known case for Bessel functions (see [8, Chap. 11, (10.23)]). For other fixed values of α (real or complex, subject to $|\arg(\alpha+1)|\le\pi-\delta$ and $\alpha\ne-1$) we can utilize a symbolic algebra program (such as MAPLE) to expand the right-hand side, via Stirling's formula, in inverse powers of u, thereby determining the coefficients l_s. For example, we find that the first two are given by

$$l_1=-\frac{\alpha}{48(1+\alpha)} \tag{4.33}$$

and

$$l_2=\frac{\alpha(6048+6048\alpha+2021\alpha^2)}{1658880(1+\alpha)^3}. \tag{4.34}$$

Let us summarize our main result.

THEOREM 4.1. *Denote new parameters by*

$$u=n+\frac{1}{2},\quad \alpha=\frac{a-2}{u}, \tag{4.35}$$

let turning points $z_{1,2}$ be defined by (3.7)*, and define a new dependent variable ζ by* (3.16)*. Then if $\alpha=O(1)$, $|\alpha+1|\ge\delta>0$, and $|\alpha+2|\ge\delta>0$, the reverse generalized Bessel polynomial has the uniform asymptotic expansion*

$$\theta_n(uz;a)=\frac{e^{-u(1+\alpha)\pi i/2}\sqrt{\pi}2^{1+u\alpha/2}u^{u-1/3}e^{u\alpha/2}}{(1+\alpha)^{u(1+\alpha)/2}}\left[1+\sum_{s=0}^{N-1}\frac{l_s}{u^{2s+1}}\right]^{-1} \tag{4.36}$$
$$\times\left\{\frac{\zeta}{(z-z_1)(z-z_2)}\right\}^{1/4}z^{u+u\alpha/2}e^{uz}W_{2N+1,0}(u,\zeta),$$

where the constants $\{l_s\}_{s=0}^{\infty}$ are implicitly defined via the asymptotic expansion (4.32)*. In the expansion* (4.36)

$$W_{2N+1,0}(u,\zeta)=\mathrm{Ai}(u^{2/3}\zeta)\sum_{s=0}^{N}\frac{A_s(\zeta)}{u^{2s}}+\frac{1}{u^2}\frac{\mathrm{d}\ \mathrm{Ai}(u^{2/3}\zeta)}{\mathrm{d}\zeta}\sum_{s=0}^{N-1}\frac{B_s(\zeta)}{u^{2s}}+\varepsilon_{2N+1,0}(u,\zeta), \tag{4.37}$$

where the coefficients $A_s(\zeta)$ *and* $B_s(\zeta)$ *are defined by* (4.3) *and* (4.4), *and the error term* $\varepsilon_{2N+1,0}(u,\zeta)$ *is bounded by* (4.7), *uniformly for* $\zeta \in \mathbf{Z}_0(\alpha)$ *(this domain of validity being described above). The corresponding region of validity in the* z *plane, for the case* α *real and lying in the interval* $-1 < \alpha < \infty$, *consists of all points in the cut plane depicted in Figure* 3, *with the exception of a neighborhood of* $z = z_2$ *(which is a singularity of the* $z - \zeta$ *transformation). In particular, when* $-1 < \alpha < \infty$ *the asymptotic expansion* (4.36) *and* (4.37) *is uniformly valid for* $\mathrm{Im}(z) \geq 0$.

Remark. For a corresponding asymptotic expansion which is valid in the lower half plane (and in particular in a neighborhood of the turning point at $z = z_2(\alpha)$) one can use (4.36) and (4.37) with α replaced by $\overline{\alpha}$ (so that ζ is suitably redefined), and the reflection formula

$$\theta_n(z;a) = \overline{\theta_n(\bar{z};\bar{a})}. \tag{4.38}$$

In deriving (4.36) we used the behavior of both sides as $z \to \infty$. Therefore, as a powerful check on the correctness of this approximation (which is supposed to be valid at the singularity at $z = 0$), we compare both sides as $uz \to 0$ (with $u \to \infty$). Now since $u^{2/3}\zeta \to \infty e^{-2\pi i/3}$ as $z \to 0$ and $u \to \infty$, we use (3.7), (3.22), (4.14), and (4.37) to find that in this case the right-hand side of (4.36) (for the case $N = 0$) takes the form

$$\sqrt{\frac{2}{u(2+\alpha)}}\left(\frac{2+\alpha}{1+\alpha}\right)^{u\alpha}\left\{\frac{u(2+\alpha)^2}{2e(1+\alpha)}\right\}^u \{1+O(u^{-1})\}\{1+O(uz)\}. \tag{4.39}$$

As a comparison, from (1.11) we observe that

$$\theta_n(uz;a) = \frac{\Gamma(2n+a-1)}{2^n\Gamma(n+a-1)}\{1+O(uz)\} \tag{4.40}$$

as $uz \to 0$. Now from (4.35), and with the aid of Stirling's formula (with $u \to \infty$ and α fixed), it follows that

$$\begin{aligned}\frac{\Gamma(2n+a-1)}{2^n\Gamma(n+a-1)} &= \frac{\Gamma(2u+u\alpha)}{2^{u-1/2}\Gamma(u+u\alpha+\frac{1}{2})} \\ &= \sqrt{\frac{2}{u(2+\alpha)}}\left(\frac{2+\alpha}{1+\alpha}\right)^{u\alpha}\left\{\frac{u(2+\alpha)^2}{2e(1+\alpha)}\right\}^u \{1+O(u^{-1})\}.\end{aligned} \tag{4.41}$$

Thus, we have verified that (4.39) and (4.40) are indeed equivalent.

5. Expansions involving elementary functions: Complex *z*. We now obtain asymptotic expansions for the three solutions in terms of elementary (exponential) functions. These will be even simpler than the Airy function expansions of the previous section, but they will not be valid in the neighborhood of either turning point, although they will still be valid at $z = 0$ and $z = \infty$. Moreover, the (complex) domains of validity will contain the entire real z axis. Although the results will be valid for certain complex values of α, for simplicity we shall assume in what follows that α is real and lying in the interval $-1 < \alpha < \infty$.

The relevant asymptotic theory, the Liouville–Green approximation in the complex plane, is given by Olver in [8, Chap. 10, sects. 1–5]. From (2.02) of [8] we first introduce a new independent variable by

$$\xi = \int f^{1/2}(\alpha,z)\mathrm{d}z = \int \frac{\{z^2+\alpha z+(1+\frac{1}{2}\alpha)^2\}^{1/2}}{z}\mathrm{d}z. \tag{5.1}$$

Upon integration we find that

$$\xi = Z + \frac{1}{2}(2+\alpha)\ln\left\{\frac{2(2+\alpha)Z - 2\alpha z - (2+\alpha)^2}{z}\right\} + \frac{1}{2}\alpha\ln\{2Z + 2z + \alpha\}, \tag{5.2}$$

where

$$Z = \{(z-z_1)(z-z_2)\}^{1/2} = \left\{z^2 + \alpha z + \left(1 + \frac{1}{2}\alpha\right)^2\right\}^{1/2}. \tag{5.3}$$

The branches associated with (5.2) are as specified in section 3, so that in terms of the variable ω defined by (3.18) we have

$$\xi = \omega + \frac{1}{2}(1+\alpha)\ln(1+\alpha) + \frac{1}{2}(4+3\alpha)\ln(2) + \frac{1}{2}(1+\alpha)\pi i. \tag{5.4}$$

For identification purposes we shall require knowledge of the behavior of ξ as z approaches the singularities of the differential equation (3.4). First, from (3.19) and (5.4) we see that as $\mathrm{Re}\ z \to \infty$

$$\xi = z + \frac{1}{2}\alpha\ln(z) + 2(1+\alpha)\ln(2) + \frac{1}{2}\alpha + O(z^{-1}), \tag{5.5}$$

and as $\mathrm{Re}\ z \to -\infty$

$$\xi = -z - \frac{1}{2}\alpha\ln(z) + (2+\alpha)\ln(2) + (1+\alpha)\ln(1+\alpha) - \frac{1}{2}\alpha + (1+\alpha)\pi i + O(z^{-1}). \tag{5.6}$$

Likewise, as $z \to 0$ we find that

$$\begin{aligned}\xi = {} & \frac{1}{2}(2+\alpha)\ln(z) + (1+\alpha)\ln(1+\alpha) + (3+2\alpha)\ln(2) \\ & - (2+\alpha)\ln(2+\alpha) + \frac{1}{2}\alpha + 1 + O(z).\end{aligned} \tag{5.7}$$

Next, we define a new dependent variable, say, $\tilde{W}$, in the usual manner

$$\tilde{W} = f^{1/4}w. \tag{5.8}$$

From the Liouville transformations (5.1) and (5.8) we then obtain our desired equation

$$\frac{\mathrm{d}^2\tilde{W}}{\mathrm{d}\xi^2} = \{u^2 + \tilde{\psi}(\xi)\}\tilde{W}, \tag{5.9}$$

which is of the same form as (2.01) of [8, Chap. 10]. The Schwarzian derivative this time is given by

$$\tilde{\psi}(\xi) = \frac{4ff'' - 5f'^2}{16f^3} + \frac{g}{f}, \tag{5.10}$$

and so after some calculation, and referring to (3.5) and (3.6), we find explicitly that

$$\tilde{\psi}(\xi) = \frac{z\{(4+3\alpha)(4+\alpha)z + \alpha(2+\alpha)^2 - 4z^3\}}{32\{z^2 + \alpha z + (1+\frac{1}{2}\alpha)^2\}^3}. \tag{5.11}$$

Note that $\tilde{\psi}(\xi)$ is analytic at $z = 0$, but unlike $\psi(\zeta)$ given by (3.14) above, has a singularity at the turning point $z = z_1(\alpha)$.

Next, using (2.09) of [8, Chap. 10] we recursively define the coefficients

$$\mathsf{A}_{s+1}(z) = -\frac{z}{2\{z^2 + \alpha z + (1 + \frac{1}{2}\alpha)^2\}^{1/2}}\mathsf{A}_s'(z) \tag{5.12}$$

$$+ \frac{1}{32}\int_{\infty}^{z} \frac{\alpha(2+\alpha)^2 + (3\alpha+4)(4+\alpha)t - 4t^3}{\{t^2 + \alpha t + (1+\frac{1}{2}\alpha)^2\}^{5/2}}\mathsf{A}_s(t)\mathrm{d}t \quad (s = 0, 1, 2, \ldots)$$

with $\mathsf{A}_0(z) = 1$. We now can apply Theorem 3.1 of Olver [8, Chap. 10] to our equation (5.9). We express the first asymptotic solution, given by (3.02) of Olver's theorem, in the form

$$W_N^{(-1)}(u,\xi) = e^{u\xi}\sum_{s=0}^{N-1}\frac{\mathsf{A}_s(z)}{u^s} + \varepsilon_N^{(-1)}(u,\xi). \tag{5.13}$$

In order to bound the error term we specify a reference point, and the natural choice is $\xi = -\infty$: this corresponds to $z = 0$ (see (5.4), and Figures 1, 2(a), and 2(b)). Setting $j = 1$ in (3.04) of [8, Chap. 10] we obtain the bounds

$$|\varepsilon_N^{(-1)}(u,\xi)|, \left|\frac{\partial\varepsilon_N^{(-1)}(u,\xi)}{u\partial\xi}\right| \le 2|e^{u\xi}|\exp\left\{\frac{2}{u}\mathcal{V}_{\mathcal{L}^{(-1)}}(\mathsf{A}_1)\right\}\frac{\mathcal{V}_{\mathcal{L}^{(-1)}}(\mathsf{A}_N)}{u^N}. \tag{5.14}$$

The path $\mathcal{L}^{(-1)}$ connects $-\infty$ to ξ and must meet the following two requirements:

(i) it consists of a finite chain of R_2 arcs, and

(ii) as ν runs from $-\infty$ to ξ along this path, Re ν is nondecreasing.

The domain of validity of the expansion consists of all points that can be accessed by such a path. For the case α real and lying in the interval $-1 < \alpha < \infty$ this domain is depicted in Figure 7(a).

The second solution, given by (3.03) of Theorem 3.1 of [8, Chap. 10], actually supplies two more solutions. One is recessive at $z = +\infty$ and the other at $z = -\infty$. These are of the same form, but the reference point is different for each: one is at $\xi = +\infty$ corresponding to $z = -\infty$ (see Figure 2(a)), and the other is at $\xi = +\infty$ corresponding to $z = +\infty$ (see Figure 2(b)). Using the suffix $j = 1$ for the former, and $j = 0$ for the latter, we thus have the following two independent asymptotic solutions:

$$W_N^{(j)}(u,\xi) = e^{-u\xi}\sum_{s=0}^{N-1}(-1)^s\frac{\mathsf{A}_s(z)}{u^s} + \varepsilon_N^{(j)}(u,\xi), \tag{5.15}$$

where

$$|\varepsilon_N^{(j)}(u,\xi)|, \left|\frac{\partial\varepsilon_N^{(j)}(u,\xi)}{u\partial\xi}\right| \le 2|e^{-u\xi}|\exp\left\{\frac{2}{u}\mathcal{V}_{\mathcal{L}^{(j)}}(\mathsf{A}_1)\right\}\frac{\mathcal{V}_{\mathcal{L}^{(j)}}(\mathsf{A}_N)}{u^N}. \tag{5.16}$$

The paths of integration run from the respective reference points at infinity to the point ξ, and this time Re ν must be nonincreasing as ν runs from ∞ to the point ξ. The z domains of validity of the bounds (5.16) are depicted in Figures 7(b) ($j = 0$) and 7(c) ($j = 1$) (for the case α real and lying in the interval $-1 < \alpha < \infty$).

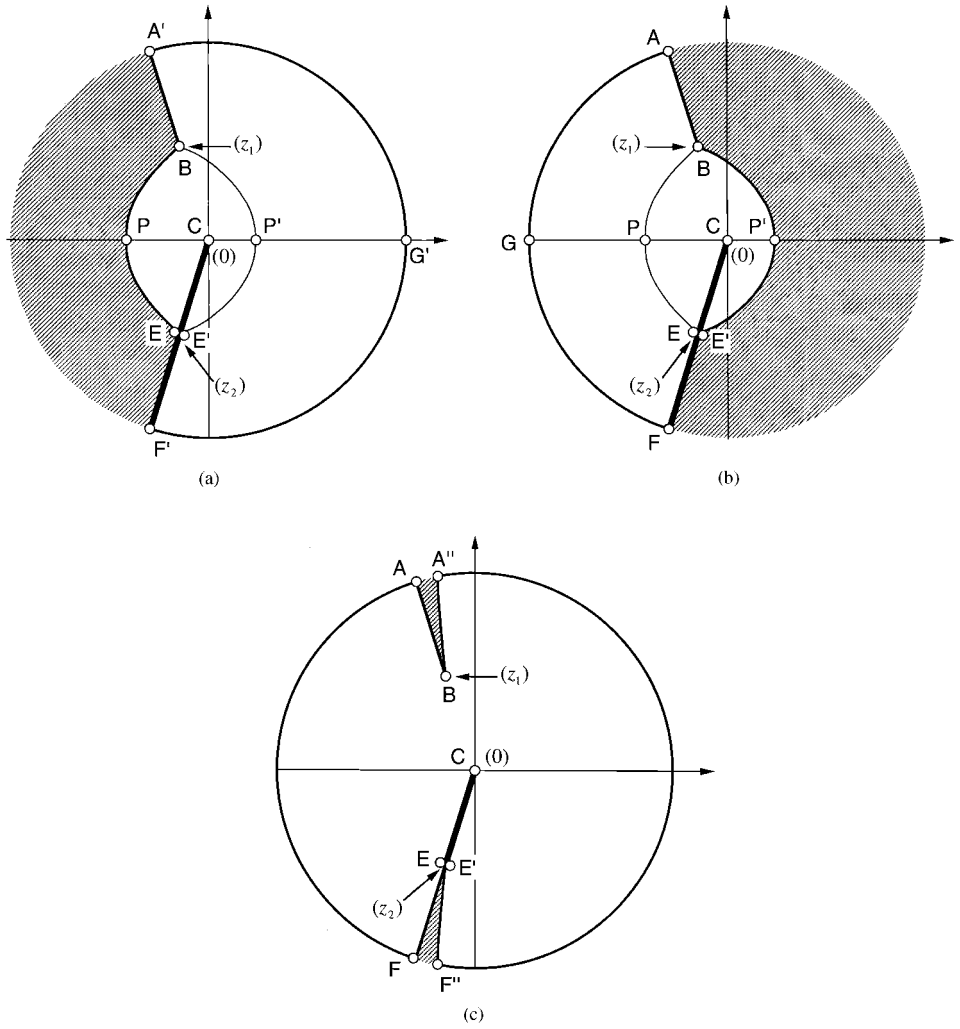


FIG. 7. *z plane.*

We now identify the asymptotic solutions with the standard solutions in the manner of section 4. On employing (5.5) we find that

$$\theta_n(uz;a) = \frac{4^{u(1+\alpha)}u^{u-1/2}e^{u\alpha/2}z^{u(2+\alpha)/2}e^{uz}}{\{z^2+\alpha z+(1+\frac{1}{2}\alpha)^2\}^{1/4}}W_N^{(0)}(u,\xi), \tag{5.17}$$

both being solutions which are recessive at $z=+\infty$.

Next, we identify solutions which are recessive at $z=0$, and using (5.7) we find that

$$w_n^{(-1)}(uz;a) = \frac{(2+\alpha)^{2u+u\alpha+1/2}u^{u+u\alpha/2+1/2}}{2^{3u+2u\alpha+1/2}e^{u(2+\alpha)/2}(1+\alpha)^{u(1+\alpha)}\Gamma(2u+u\alpha+1)} \tag{5.18}$$

$$\times\left\{1+\sum_{s=1}^{N-1}\frac{\mathsf{A}_s(0)}{u^s}\right\}^{-1}\frac{z^{1/2}}{\{z^2+\alpha z+(1+\frac{1}{2}\alpha)^2\}^{1/4}}W_N^{(-1)}(u,\xi).$$

Finally, for the solutions recessive at $z=-\infty$ we use (5.6) to obtain

$$w_n^{(1)}(uz;a)=(-1)^n e^{u\alpha\pi i}2^{u+u\alpha-1/2}e^{-u\alpha/2}u^{u\alpha/2}(1+\alpha)^{u+u\alpha} \tag{5.19}$$

$$\times\left\{1+\sum_{s=1}^{N-1}(-1)^s\frac{\mathsf{A}_s(-\infty)}{u^s}\right\}^{-1}\frac{z^{1/2}}{\{z^2+\alpha z+(1+\frac{1}{2}\alpha)^2\}^{1/4}}W_N^{(1)}(u,\xi),$$

where

$$\mathsf{A}_s(-\infty)=\lim_{z\to-\infty}\mathsf{A}_s(z). \tag{5.20}$$

A compound asymptotic expansion for $\theta_n(uz;a)$, which is valid in a part of the left half plane (as depicted in Figure 7(b) when $-1<\alpha<\infty$), is easily obtained by inserting the expansions (5.18) and (5.19) into the connection formula (2.13) (and referring to (2.3)). Such an expansion is valid along the curve PBE (except near the turning points at the ends of this curve). The significance of this is that the zeros of $\theta_n(uz;a)$ are located near this curve when n is large.

In the next two sections we consider the case z real, and we also give a particularly simple procedure for calculating the coefficients $\{\mathsf{A}_s(z)\}_{s=1}^{\infty}$, as well as the limiting constants given by (5.20).

6. Expansions involving elementary functions: Positive *z*. We continue to assume that $-1<\alpha<\infty$ but now restrict z to $0<z<\infty$. Let us first introduce the positive parameter

$$p=1/Z=\left\{z^2+\alpha z+\left(1+\frac{1}{2}\alpha\right)^2\right\}^{-1/2}. \tag{6.1}$$

Then, as z decreases from ∞ to $-\infty$, p increases from 0 to a maximum value of $\{1+\alpha\}^{-1/2}$ (corresponding to $z=-\frac{1}{2}\alpha$), and then decreases back to 0. In terms of this variable we can express the recursion formula (5.12) for the coefficients in the form

$$\mathsf{A}_{s+1}(z(p))=\frac{1}{4}p^2[2-2(1+\alpha)p^2-\alpha p\{1-(1+\alpha)p^2\}^{1/2}]\frac{\mathrm{d}}{\mathrm{d}p}\mathsf{A}_s(z(p)) \tag{6.2}$$

$$+\frac{1}{16}\int_0^p\left[\frac{\alpha q\{5(1+\alpha)q^2-3\}}{\{1-(1+\alpha)q^2\}^{1/2}}-10(1+\alpha)q^2+2\right]\mathsf{A}_s(z(q))\mathrm{d}q.$$

A further simplification in this comes from introducing a variable t by

$$t=\sin^{-1}(\sqrt{1+\alpha}p). \tag{6.3}$$

The branch of the inverse sine is chosen so that t depends continuously on z for $-\infty<z<\infty$, with $t=0$ corresponding to $z=\infty$, $t=\frac{1}{2}\pi$ corresponding to $z=-\frac{1}{2}\alpha$ (i.e., $p=1/\sqrt{1+\alpha}$), and $t=\pi$ corresponding to $z=-\infty$.

Remark. t is real for $-1<\alpha<\infty$ and $-\infty<z<\infty$. To see this we use (6.1) and (6.3) to obtain

$$\sin^2(t)=(1+\alpha)p^2=\frac{1+\alpha}{1+\alpha+(z+\frac{1}{2}\alpha)^2}. \tag{6.4}$$

Thus,

$$0 < \sin(t) \leq 1 \tag{6.5}$$

with $\sin(t) \to 0$ as $z \to \pm\infty$, and $t = t_0$ when $z = 0$, where

$$t_0 = \cos^{-1}\left\{\frac{\alpha}{2+\alpha}\right\}, \tag{6.6}$$

the branch of the inverse cosine being such that $0 < t_0 < \pi$ when $-1 < \alpha < \infty$. In summary, for fixed α in $(-1, \infty)$, the z interval $-\infty < z < \infty$ is mapped one to one by (6.3) to the t interval $0 < t < \pi$.

We now consider the coefficients defined by (6.2) as functions of t. If we write

$$\mathsf{A}_s(z) = V_s(t), \tag{6.7}$$

we obtain the recursion formula

$$V_{s+1}(t) = \frac{1}{4}\sin^2(t)\left[\frac{2\cos(t)}{\sqrt{1+\alpha}} - \frac{\alpha\sin(t)}{1+\alpha}\right]\frac{\mathrm{d}}{\mathrm{d}t}V_s(t) \tag{6.8}$$

$$+\frac{1}{16}\int_0^t \left[\frac{2\cos(\tau)\{1-5\sin^2(\tau)\}}{\sqrt{1+\alpha}} + \frac{\alpha\sin(\tau)\{5\sin^2(\tau)-3\}}{1+\alpha}\right] V_s(\tau)\mathrm{d}\tau \quad (s = 0, 1, 2, \ldots)$$

with $V_0(t) = 1$. Note that the lower limit $t = 0$ in this integral corresponds to $z = \infty$, but *not* to $z = -\infty$. In other words, each coefficient vanishes as $z \to \infty$ but not necessarily as $z \to -\infty$. In fact, for $z = -\infty$ the coefficients are given by

(6.9)

$$\begin{aligned}\mathsf{A}_s(-\infty) &= V_s(\pi)\\ &= \frac{1}{16}\int_0^\pi \left[\frac{2\cos(\tau)\{1-5\sin^2(\tau)\}}{\sqrt{1+\alpha}} + \frac{\alpha\sin(\tau)\{5\sin^2(\tau)-3\}}{1+\alpha}\right] V_{s-1}(\tau)\mathrm{d}\tau.\end{aligned}$$

From (6.8) we observe that each $V_s(t)$ is a multivariate polynomial of the two variables $\cos(t)$ and $\sin(t)$. The recursion (6.8) is therefore easily handled by a symbolic algebra program, and, for example, we find that

$$V_1(t) = \frac{\sin(t)(5\sin^2(t)-3)}{24\sqrt{1+\alpha}} + \frac{\alpha(\cos(t)-1)(5\cos^2(t)+5\cos(t)-1)}{48(1+\alpha)} \tag{6.10}$$

and

(6.11)

$$\begin{aligned}V_2(t) = {}& \frac{\sin^2(t)}{1152(1+\alpha)}(81 - 462\sin^2(t) + 385\sin^4(t)) - \frac{\alpha^2(\cos(t)-1)^2 P(\cos(t))}{4608(1+\alpha)^2}\\ &+ \frac{\alpha\sin(t)(\cos(t)-1)(385\cos^4(t)+385\cos^3(t)-231\cos^2(t)-226\cos(t)+2)}{1152(1+\alpha)^{3/2}},\end{aligned}$$

where

$$\begin{aligned}P(x) = {}& 145x^{10} + 290x^9 - 435x^8 - 1160x^7 + 290x^6\\ &+ 1740x^5 + 50x^4 - 1640x^3 - 666x^2 + 298x + 143.\end{aligned} \tag{6.12}$$

The Liouville transformation, for $0 < z < \infty$, is the same as in section 5, and in terms of t we find that

$$\xi = \frac{\sqrt{1+\alpha}}{\sin(t)} - \frac{\alpha}{2}\ln\left\{\tan\left(\frac{1}{2}t\right)\right\} - (2+\alpha)\tanh^{-1}\left\{\frac{2\sqrt{1+\alpha}\tan(\frac{1}{2}t)+\alpha}{2+\alpha}\right\}. \tag{6.13}$$

The interval $0 < z < \infty$ is mapped to $-\infty < \xi < \infty$, with $\xi = -\infty$ corresponding to $z = 0$, and $\xi = \infty$ corresponding to $z = \infty$.

The identification of solutions (recessive at $z = \infty$ and $z = 0$) proceeds as before, and we merely state the results:

(6.14)

$$\theta_n(uz;a) = 4^{u(1+\alpha)}u^{u-1/2}e^{u\alpha/2}p^{1/2}z^{u(2+\alpha)/2}e^{uz-u\xi}\left\{\sum_{s=0}^{N-1}(-1)^s\frac{V_s(t)}{u^s} + \eta_N^-(u,t)\right\},$$

where

$$|\eta_N^-(u,t)| \le 2\exp\left\{\frac{2}{u}\mathcal{V}_{0,t}(V_1)\right\}\frac{\mathcal{V}_{0,t}(V_N)}{u^N} \tag{6.15}$$

uniformly for $0 < z < \infty$, and

(6.16)

$$\mathbf{M}(n+a-1,2n+a,2uz) = \frac{(2+\alpha)^{2u+u\alpha+1/2}}{2^{3u+2u\alpha+1/2}e^{u(2+\alpha)/2}(1+\alpha)^{u(1+\alpha)}\Gamma(2u+u\alpha+1)}$$
$$\times \left\{1+\sum_{s=1}^{N-1}\frac{V_s(t_0)}{u^s}\right\}^{-1} p^{1/2}z^{-u(2+\alpha)/2}e^{uz+u\xi}\left\{\sum_{s=0}^{N-1}\frac{V_s(t)}{u^s} + \eta_N^+(u,t)\right\},$$

where

$$|\eta_N^+(u,t)| \le 2\exp\left\{\frac{2}{u}\mathcal{V}_{t,t_0}(V_1)\right\}\frac{\mathcal{V}_{t,t_0}(V_N)}{u^N} \tag{6.17}$$

uniformly for $0 < z < \infty$.

7. Expansions involving elementary functions: Negative *z*. The variable ξ is not real when $-\infty < z < 0$, so we introduce the following real variable instead:

$$\tilde{\xi} = Z + \frac{1}{2}(2+\alpha)\ln\left\{\frac{2(2+\alpha)Z - 2\alpha z - (2+\alpha)^2}{|z|}\right\} + \frac{1}{2}\alpha\ln\{2Z+2z+\alpha\}. \tag{7.1}$$

When $-\infty < z < 0$ this differs from ξ, as given by (5.2) (with $\arg(z) = \pi$), by a constant: that is,

$$\tilde{\xi} = \xi - \frac{1}{2}(2+\alpha)\pi i. \tag{7.2}$$

Hence, the transformed equation is still of the form (5.9) (with ξ replaced by $\tilde{\xi}$). In order to identify recessive solutions we require the following asymptotic forms:

$$\tilde{\xi} = \frac{1}{2}(2+\alpha)\ln|z| + (1+\alpha)\ln(1+\alpha) + (3+2\alpha)\ln(2) \tag{7.3}$$
$$- (2+\alpha)\ln(2+\alpha) + \frac{1}{2}\alpha + 1 + O(z)$$

as $z \to 0$ and

$$\tilde{\xi} = |z| - \frac{1}{2}\alpha \ln|z| + (2+\alpha)\ln(2) + (1+\alpha)\ln(1+\alpha) - \frac{1}{2}\alpha + O(z^{-1}) \tag{7.4}$$

as $z \to -\infty$.

Now for z negative a real solution of (3.4) which is recessive at $z = -\infty$ is given by $U(n+1, 2n+a, 2u|z|)$, where the confluent hypergeometric function U is defined by (1.5). The behavior of this function at $z = -\infty$ follows from [8, Chap. 7, (10.01)], and using this and (7.4) we arrive at the asymptotic expansion

$$U(n+1, 2n+a, 2u|z|) = 2^{u+u\alpha-1/2}(1+\alpha)^{u(1+\alpha)}u^{-u-1/2}e^{-u\alpha/2} \tag{7.5}$$
$$\times \left\{1 + \sum_{s=1}^{N-1}(-1)^s \frac{V_s(\pi)}{u^s}\right\}^{-1} p^{1/2}|z|^{-u(2+\alpha)/2}e^{u|z|-u\tilde{\xi}} \left\{\sum_{s=0}^{N-1}(-1)^s \frac{V_s(t)}{u^s} + \tilde{\eta}_N^-(u,t)\right\}.$$

Here $\tilde{\eta}_N^-(u,t)$ is bounded by

$$|\tilde{\eta}_N^-(u,t)| \le 2\exp\left\{\frac{2}{u}\mathcal{V}_{t,\pi}(V_1)\right\}\frac{\mathcal{V}_{t,\pi}(V_N)}{u^N} \tag{7.6}$$

uniformly for $t_0 < t < \pi$ (i.e., $-\infty < z < 0$).

The corresponding expansion for the solution recessive at $z = 0$ is found to be

$$\mathbf{M}(n+a-1, 2n+a, 2uz) = \frac{(2+\alpha)^{2u+u\alpha+1/2}}{2^{3u+2u\alpha+1/2}e^{u(2+\alpha)/2}(1+\alpha)^{u(1+\alpha)}\Gamma(2u+u\alpha+1)} \tag{7.7}$$
$$\times \left\{1 + \sum_{s=1}^{N-1}\frac{V_s(t_0)}{u^s}\right\}^{-1} p^{1/2}|z|^{-u(2+\alpha)/2}e^{uz+u\tilde{\xi}} \left\{\sum_{s=0}^{N-1}\frac{V_s(t)}{u^s} + \tilde{\eta}_N^+(u,t)\right\}.$$

The bound on $\tilde{\eta}_N^+(u,t)$ is the same as that given by (6.17) for $\eta_N^+(u,t)$.

The Bessel polynomial $\theta_n(uz;a)$ is of most interest, but it is not recessive in $-\infty < z < 0$, and hence cannot be identified directly with a Liouville–Green asymptotic solution. However, we obtain a compound asymptotic expansion which is uniformly valid in this interval via the connection formula

$$\theta_n(uz;a) = 2^{n+a-1}u^{2n+a-1}n!|z|^{2n+a-1} \tag{7.8}$$
$$\times \left[\frac{e^{2uz}U(n+1, 2n+a, 2u|z|)}{\Gamma(n+a-1)} + (-1)^n\mathbf{M}(n+a-1, 2n+a, 2uz)\right].$$

Thus, we substitute (7.5) and (7.7) into this to obtain the uniform asymptotic expansion

$$\theta_n(uz;a) = n!e^{-u\alpha/2}p^{1/2}|z|^{u(2+\alpha)/2}e^{uz} \tag{7.9}$$
$$\times \left[C_N^-e^{-u\tilde{\xi}}\left\{\sum_{s=0}^{N-1}(-1)^s\frac{V_s(t)}{u^s} + \tilde{\eta}_N^-(u,t)\right\} + (-1)^nC_N^+e^{u\tilde{\xi}}\left\{\sum_{s=0}^{N-1}\frac{V_s(t)}{u^s} + \tilde{\eta}_N^+(u,t)\right\}\right],$$

where

$$C_N^- = \frac{2^{2u(1+\alpha)}(1+\alpha)^{u(1+\alpha)}u^{u+u\alpha-1/2}}{\Gamma(u+u\alpha+\frac{1}{2})}\left\{1+\sum_{s=1}^{N-1}(-1)^s\frac{V_s(\pi)}{u^s}\right\}^{-1} \tag{7.10}$$

and

$$C_N^+ = \frac{(2+\alpha)^{2u+u\alpha+1/2}u^{u(2+\alpha)}}{2^{u(2+\alpha)}(1+\alpha)^{u(1+\alpha)}e^u\Gamma(2u+u\alpha+1)}\left\{1+\sum_{s=1}^{N-1}\frac{V_s(t_0)}{u^s}\right\}^{-1}. \tag{7.11}$$

REFERENCES

[1] S. Bochner, *Über Sturm-Liouvillsche Polynomsysteme*, Math. Z., 29 (1929), pp. 730–736.

[2] W. G. C. Boyd and T. M. Dunster, *Uniform asymptotic solutions of a class of second-order linear differential equations having a turning point and a regular singularity, with an application to Legendre functions*, SIAM J. Math. Anal., 17 (1986), pp. 422–450.

[3] A. J. Carpenter, *Asymptotics for the zeros of the generalized Bessel polynomials*, Numer. Math., 62 (1992), pp. 465–482.

[4] C. Chester, B. Friedman, and F. Ursell, *An extension of the method of steepest descents*, Proc. Cambridge Philos. Soc., 53 (1957), pp. 599–611.

[5] M. G. de Bruin, E. B. Saff, and R. S. Varga, *On the zeros of the generalized Bessel polynomials*, Indag. Math., 43 (1981), pp. 1–25.

[6] E. Grosswald, *Bessel Polynomials*, Lecture Notes in Math. 698, Springer, New York, 1978.

[7] H. L. Krall and O. Frink, *A new class of orthogonal polynomials: The Bessel polynomial*, Trans. Amer. Math. Soc., 65 (1949), pp. 100–115.

[8] F. W. J. Olver, *Asymptotics and Special Functions*, Academic Press, New York, 1974. Reprinted by AK Peters, Wellesley, 1997.

[9] F. W. J. Olver, *Second-order linear differential equations with two turning points*, Philos. Trans. Roy. Soc. London Ser. A, 278 (1975), pp. 137–174.

[10] V. Romanovsky, *Sur quelques classes nouvelles des polynômes orthogonaux*, C. R. Acad. Sci. Paris Ser. I Math., 188 (1929), pp. 1023–1025.

[11] W. E. Thompson, *Delay network having maximally flat frequency characteristics*, Proc. Institute Electr. Engineers, 96 (1949), p. 487.

[12] R. Wong and J.-M. Zhang, *Asymptotic expansions of the generalized Bessel polynomials*, J. Comput. Appl. Math., 85 (1997), pp. 87–112.

SIAM J. MATH. ANAL.
Vol. 32, No. 5, pp. 1014–1031

EXISTENCE OF LARGE AMPLITUDE PERIODIC WAVES IN TWO-FLUID FLOWS OF INFINITE DEPTH*

S. M. SUN†

Abstract. Two-dimensional periodic traveling gravity waves in a two-fluid flow are considered, where the flow has no rigid boundaries. Each fluid is inviscid, incompressible, and irrotational and the density ratio of the upper fluid to the lower fluid is between zero and one. The governing equations are first transformed into a single nonlinear integral equation using the Hilbert transform and the corresponding integral operator is compact in certain Banach spaces after a cut-off function is introduced. By a global bifurcation theorem, it is shown that there exist periodic waves of large amplitude on the interface until either the bifurcation parameter goes to infinity or the function of the wave profile and its first-order derivative are not in the classical Hölder space. It is also noted that the nonlinear integral equation is very general and can be used to study the waves of large amplitude numerically.

Key words. large amplitude wave, two-fluid flow

AMS subject classifications. 35Q35, 76B15

PII. S0036141099352728

1. Introduction.

Mathematical investigation of two-dimensional propagating gravity waves in fluids has attracted a great deal of attention in the last several decades. The fluids are assumed to be inviscid, incompressible, and irrotational and waves are moving with a uniform speed in the fluids without changing its form. There are two types of waves that we consider, single-hump waves, called solitary waves, and periodic waves. Although such waves have been observed in experiments and derived formally from the exact equations using the asymptotic method, it is of interest to study these waves mathematically and show rigorously the existence of such waves from the exact nonlinear governing equations.

For waves in a one-layered fluid with free surface, there are many theoretical works in the literature focusing on the existence of periodic and solitary waves of small amplitude (for example, Levi-Civita [1], Struik [2], Nekrasov [3], Friedrichs and Hyers [4], and Beale [5]). Krasovskii [6] first studied existence of large amplitude periodic waves using a similar integral formulation by Nekrasov [3] and found a global existence of periodic waves. Keady and Norbury [7] extended the Krasovskii set of periodic waves using a global bifurcation theorem. Amick and Toland [8] proved existence of solitary waves of all amplitudes from zero up to and including that of solitary wave of greatest height. Further study on the largest amplitude solution can be found in [9, 10, 11] and the references therein. Recently, Buffoni, Dancer, and Toland [12] gave a significant new contribution on this problem and showed a very complicated structure of secondary bifurcation near the solution of largest amplitude using a variational method.

For interfacial waves between two fluids of different densities without boundaries, Holyer [13] performed extensive numerical computations and found interfacial waves

*Received by the editors March 1, 1999; accepted for publication (in revised form) July 21, 2000; published electronically January 16, 2001. This research was partly supported by National Science Foundation grants DMS-9623060 and DMS-9971764.
http://www.siam.org/journals/sima/32-5/35272.html
†Department of Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061 (sun@math.vt.edu).

for which the interface profile is vertical at some points if amplitude of waves becomes large. Meiron and Saffman [14] computed the solutions again and established the existence of overhanging waves, for which some portions of the heavier fluid lie above the lighter fluid. Subsequently, Turner and VandenBroeck [15] and Grimshaw and Pullin [16] carried out extensive numerical computations and found that the situations are quite complicated. The large amplitude waves can oscillate alternating between an overhanging wave and a single-valued wave indefinitely [15] or approach to a mushroom-like limiting configuration [16]. Similar wave profiles can be calculated for interfacial waves in two-fluid flows of finite depth [17, 18], while Amick and Turner [19] proved the existence of large-amplitude waves with wave profiles having no vertical tangent lines. By contrast, there is almost no theoretical work to show the existence of large amplitude periodic waves in two-layered fluids of infinite depth. To study overhanging waves, a formulation similar to the one by Nekrasov [3] for one-layered fluids must be obtained, which was an open problem and given to the author by Benjamin [20].

In this paper, we shall derive an integral formulation for two-dimensional periodic traveling gravity waves in two fluids without boundaries and then give a mathematical proof of the existence of such waves of small and large amplitude using the global bifurcation theory. The result can be stated as follows. Let the wave be symmetric with respect to its crest and periodic. The interface of two fluids is determined by Φ, which measures the angle between the tangent line of the interfacial curve and horizontal direction. It is shown that if the wave speed is near its first critical value, nonuniform waves of small amplitude will bifurcate from the uniform state and the solutions in this branch always satisfy $\Phi \in [0, \pi]$. When amplitude of the waves in this branch becomes large, the bifurcation branch can be extended infinitely until either the bifurcation parameter goes to infinity or the function Φ and its first-order derivative are no longer Hölder continuous. Our formulation of the problem is quite general and can also be used for numerical calculations of the wave profiles.

The idea to prove the result can be summarized as follows. The exact equations are first transformed into several integral equations using the Hilbert transform. Then some tricky algebraic manipulations are used to transform the integral equations into one integral equation of Φ so that the corresponding integral operator is compact in a Banach space under some restrictions on Φ. To have an equation valid for all functions Φ, the integral equation is further transformed into another integral equation using a cut-off function. Therefore, finding the solutions of the exact governing equations becomes finding fixed points of a compact operator. Then we view the problem as a global bifurcation problem with a bifurcation parameter related to the amplitude of the wave, which was first proposed by Keady and Norbury [7] for one-layered fluids. By a global bifurcation theorem [21, 22], the existence of global bifurcation branch of solutions is obtained. Finally, using some properties of solutions derived from governing equations, we show that the solution branch cannot be bounded. Therefore, the bifurcation branch must be extended to infinity. A similar idea can be applied to derive an integral equation for a two-fluid system with either one or two horizontal rigid boundaries.

The paper is organized as follows. In section 2, the governing equations are given and some properties of the solutions are stated. Section 3 gives the integral representation of solutions for the problem. In section 4, the integral equation is modified using a cut-off function so that the equation is valid for any functions without any artificial restriction. In section 5, the global bifurcation theorem is used to obtain

a global branch of solutions for the problem. The proofs of some properties of solutions and compactness of the operators are given in Appendices A and B, respectively.

2. Formulation of the problem. Consider a two-dimensional periodic progressing wave moving with a constant speed $-c$ at the interface of two fluids. In reference to a rectangular coordinate (x, y) system moving with the wave, the flow is steady, where the x-axis is the horizontal direction with the velocity $(c, 0)$ as the positive direction and the y-axis is the vertical direction upwards through the crest of the wave so that the gravity acts in the negative y direction. The interface is symmetric with respect to the y-axis and the period of the wave is τ so that the tangent lines of the interface at $x = -\tau/2$ and $x = \tau/2$ are horizontal. In the following, we use f^+ and f^- to denote a quantity f in upper and lower fluids, respectively. The densities of the fluids are ρ^+ and ρ^- with $\rho^- > \rho^+$.

Since the fluids are incompressible and irrotational, the velocity potentials $\varphi^\pm$ and stream functions $\psi^\pm$ are well defined and the velocity vectors are given by $(\varphi^\pm_x, \varphi^\pm_y)$. It is well known that the functions $w^\pm(x, y) = w^\pm(z) = \varphi^\pm(x, y) + i\psi^\pm(x, y)$ are analytic in the upper and lower fluids with respect to $z = x + iy$. Without loss of generality, we choose $\varphi^\pm = 0$ at the crest $x = 0$ and $\psi^\pm = 0$ at the interface. In addition, there are Bernoulli equations at the interface,

$$(p^\pm/\rho^\pm) + (1/2)((u^\pm)^2 + (v^\pm)^2) + gy^\pm = H^\pm\,,$$

where $p(x, y)$ is the pressure, g is the gravitational acceleration constant, $H^\pm$ are constants called the Bernoulli heads, and $y^+ = y^-$ at the interface. By choosing the origin of the y-axis so that $H^\pm = c^2/2$, the continuity of the pressure across the interface is

$$\begin{aligned}\rho^+((1/2)((u^+)^2 + (v^+)^2) + gy^+ - (c^2/2)) \\ = \rho^-((1/2)((u^-)^2 + (v^-)^2) + gy^- - (c^2/2))\end{aligned} \tag{1}$$

at $\psi^\pm = 0$. Here, all functions are periodic in x with period τ and the choice of the origin for the y-axis makes the interface of undisturbed state at $y = 0$.

Because the form of the interface is unknown, we resort to conformal mappings to overcome this difficulty. We try to find a conformal mapping which maps two infinite regions in the z-plane occupied by two fluids onto two regions in another complex plane $U = \xi + i\eta = re^{i\theta}$ with the unit circle deleted so that the interface is mapped to the circumference of the unit circle. The upper fluid is mapped onto $r = |U| < 1$ with the infinity corresponding to the center of the circle and the lower fluid is mapped onto $r > 1$. By the symmetry and periodicity of the interface, two points at $x = -\tau/2$ and $x = \tau/2$ on the interface are mapped to $\theta = -\pi, \pi$, and $x = 0$ is mapped to $\theta = 0$. The circumference of the unit circle is parametrized by $\theta^\pm$ from $-\pi$ to π.

Now we let the conformal mapping from the z-plane to the U-plane have the following form:

$$\frac{dz}{dU} = -\frac{\tau i f^+(U)}{2\pi U}, \quad r = |U| < 1\,, \quad \frac{dz}{dU} = -\frac{\tau i f^-(U)}{2\pi U}, \quad r > 1\,, \tag{2}$$

where $f^\pm$ are analytic functions in $r < 1$ and $r > 1$, respectively, as will be determined later. More detailed discussion of (2) for single-layered fluids (i.e., $\rho^+ = 0$) can be found in [23]. Let $f^\pm(U) = R^\pm(r, \theta)e^{i\Phi^\pm(r,\theta)}$. From the symmetry and the positions

of crests and troughs of the wave, we have

$$R^{\pm}(r,\theta) = R^{\pm}(r,-\theta)\,, \quad \Phi^{\pm}(r,\theta) = -\Phi^{\pm}(r,-\theta)\,, \tag{3}$$
$$\Phi^{\pm}(r,0) = \Phi^{\pm}(r,-\pi) = 0\,.$$

For the sake of convenience, we denote $f^{\pm}(e^{i\theta}) \stackrel{\text{def}}{=} R^{\pm}(\theta)e^{i\Phi^{\pm}(\theta)}$ on $r = 1$. From (2), we can see that $\Phi^{\pm}(\theta)$ gives the angle between the x-axis and the tangent of the interface and $R^{\pm}(\theta)$ is the length of the tangent vector at the interface.

Now we find w in terms of U. From the interfacial conditions and the conditions at infinity, we have that

$$\psi = +\infty \quad \text{at } r = 0\,, \tag{4}$$
$$\psi = -\infty \quad \text{at } r = +\infty\,, \tag{5}$$
$$\psi = 0 \quad \text{at } r = 1\,, \tag{6}$$
$$u - iv = c \quad \text{at } r = 0, \infty. \tag{7}$$

Equations (4)–(6) are satisfied if

$$w = -\frac{\tau c i}{2\pi}\log U\,. \tag{8}$$

The definition of φ implies

$$u - iv = \varphi_x - i\varphi_y = \frac{dw}{dz} = \frac{dw}{dU}\frac{dU}{dz} = \frac{c}{f(U)}\,, \tag{9}$$

and $f^{+}(0) = f^{-}(\infty) = 1$ if condition (7) is satisfied. This fits the fact that $f^{\pm} = 1$ is the appropriate choice for the undisturbed flow.

If $R^{\pm}(\theta^{\pm})$ is never zero, then $\log f^{\pm} = \log R^{\pm} + i\Phi^{\pm}$ is analytic. By using the Poisson integral [25], we can have the following relation between real and imaginary parts of an analytic function on the unit circle (see the detailed discussion in [3] and [24]):

$$\Phi^{\pm}(\theta^{\pm}) = \mp\frac{1}{\pi}\int_{-\pi}^{\pi}\left(\sum_{n=1}^{+\infty}\frac{\sin n\theta^{\pm}\sin ns}{n}\right)\left(\frac{d}{ds}\log R^{\pm}(s)\right)ds\,, \tag{10}$$

where the kernel can also be written as

$$\sum_{n=1}^{+\infty}\frac{\sin n\theta\sin ns}{n} = \frac{1}{2}\log\left|\frac{\sin((s+\theta)/2)}{\sin((s-\theta)/2)}\right|\,. \tag{11}$$

Here, we note that the interface is parametrized by two parameters θ^{+} and θ^{-} and a value of z may not correspond to same $\theta^{\pm}$. By using z at the interface, we can consider θ^{-} as a function of θ^{+} on the unit circle. Therefore, the interfacial conditions are

$$z^{+}(\theta^{+}) = x^{+}(\theta^{+}) + iy^{+}(\theta^{+}) = z^{-}(\theta^{-}(\theta^{+})) = x^{-}(\theta^{-}(\theta^{+})) + iy^{-}(\theta^{-}(\theta^{+})),$$

which implies

$$\frac{dz^{-}}{d\theta^{-}} = \frac{dz^{+}}{d\theta^{-}} = \frac{dz^{+}}{d\theta^{+}}\frac{d\theta^{+}}{d\theta^{-}}\,. \tag{12}$$

To have periodic solutions in $\theta^{\pm}$, we need $\theta^+ = \theta^-$ at $\theta^- = 0$, π and $-\pi$. From (2) and (12), we have

$$(d\theta^+/d\theta^-) = (R^-/R^+)\exp(i(\Phi^- - \Phi^+))\,,$$

which gives

$$\Phi^+(\theta^+) = \Phi^-(\theta^-)\,, \qquad \frac{d\theta^-}{d\theta^+} = \frac{R^+(\theta^+)}{R^-(\theta^-)} \stackrel{\text{def}}{=} V(\theta^+)\,, \tag{13}$$

where θ^- is considered as a function of θ^+ on the circle and

$$\theta^- = \theta^-(\theta^+) = \int_0^{\theta^+} V(s)ds \stackrel{\text{def}}{=} L(\theta^+) \qquad \text{with} \quad \int_0^{-\pi} V(s)ds = -\pi\,. \tag{14}$$

By (1), (2), and (9), the Bernoulli equations can be rewritten as

$$\frac{1}{2}\left[\frac{1}{(R^-(\theta^-))^2} - \frac{\rho}{(R^+(\theta^+))^2}\right] + \frac{g(1-\rho)}{c^2}y^+(\theta^+) = \frac{1}{2}(1-\rho)\,, \tag{15}$$

where $\rho = \rho^+/\rho^- < 1$. By using (2) again and $u = e^{i\theta}$ on the circle,

$$\frac{dy^+}{d\theta^+} = \frac{\tau R^+ \sin\Phi^+}{2\pi}\,.$$

Then we take the derivative on both sides of (15) with respect to θ^+ to obtain

$$\begin{aligned}\frac{1}{2R^+(\theta^+)}&\left[\frac{\partial}{\partial\theta^+}\left(\frac{1}{(R^-(\theta^-))^2}\right) - \frac{\partial}{\partial\theta^+}\left(\frac{\rho}{(R^+(\theta^+))^2}\right)\right]\\ &+\frac{\tau g(1-\rho)}{2\pi c^2}\sin\Phi^+(\theta^+) = 0\,.\end{aligned} \tag{16}$$

We state a theorem which gives some properties of solution satisfying (3), (13), (14), and (16).

THEOREM 1. *Assume that there are analytic functions $f^{\pm}(u) = R^{\pm}(r,\theta)e^{i\Phi(r,\theta)} \neq 0$ for $|u| = r < 1$ and $|u| = r > 1$, where $R^{\pm}(\theta) = R^{\pm}(1,\theta)$ and $\Phi^{\pm}(\theta) = \Phi^{\pm}(1,\theta)$ are C^1 functions for $\theta \in [-\pi,\pi]$ satisfying (3), (13), (14), and (16). If $\Phi^+(\theta^+) = \Phi^-(\theta^-) \in [0,\pi]$ for $\theta^{\pm} \in [-\pi,0]$ with $\Phi^{\pm}(r,0) = \Phi^{\pm}(r,-\pi) = 0$ and $\Phi^-(r,\theta) \to 0$ as $r \to +\infty$, then $\Phi^{\pm}(\theta^{\pm}) \in (0,\pi)$ for $\theta^{\pm} \in (-\pi,0)$ and $\Phi^{\pm}_{\theta}(-\pi) > 0, \Phi^{\pm}_{\theta}(0) < 0$.*

The proof is quite technical and will be given in Appendix A. This theorem will be used to obtain the global existence of the solutions. Geometrically, the theorem gives us that if the variation of the angle between the horizontal direction and the tangent of the interface is in $[0,\pi]$ for the negative half of the wave, then tangent lines will never be horizontal except at the crests and troughs.

3. Integral representation of the solutions. In this section, we shall express the solutions of (10), (13), (13), (14), and (16) as solutions of some integral equations. From (13), we rewrite (16) as

$$\begin{aligned}\frac{1}{2}&\left[\frac{1}{R^+(\theta^+)}\frac{\partial}{\partial\theta^+}\left(\frac{V^2(\theta^+)}{(R^+(\theta^+))^2}\right) - \frac{2\rho}{3}\frac{\partial}{\partial\theta^+}\left(\frac{1}{(R^+(\theta^+))^3}\right)\right]\\ &+\frac{\tau g(1-\rho)}{2\pi c^2}\sin\Phi^+(\theta^+) = 0\,,\end{aligned}$$

which can be simplified to

$$\frac{1}{2}\left[\frac{2}{3}(V^2(\theta^+)-\rho)\frac{\partial}{\partial\theta^+}\left(\frac{1}{(R^+(\theta^+))^3}\right)+\frac{1}{(R^+(\theta^+))^3}\frac{\partial V^2(\theta^+)}{\partial\theta^+}\right] = -\frac{\tau g(1-\rho)}{2\pi c^2}\sin\Phi^+(\theta^+)\,. \tag{17}$$

If we assume that $V(\theta^+)$ and $R^+(\theta^+)$ are given, $(R^+(\theta^+))^{-3}$ in (17) is solved by

$$\left(\frac{1}{R^+(\theta^+)}\right)^3 = (V^2(\theta^+)-\rho)^{-3/2}I(\theta^+)\,, \tag{18}$$

where

$$I(\theta^+) \stackrel{\text{def}}{=} C\left(\frac{1}{\mu}-\int_0^{\theta^+}\sin\Phi^+(s)(V^2(s)-\rho)^{1/2}ds\right), \qquad C \stackrel{\text{def}}{=} \frac{3g(1-\rho)\tau}{2\pi c^2}\,, \qquad \lambda \stackrel{\text{def}}{=} \frac{(1-\rho)^{3/2}\mu}{3(1+\rho)}\,, \tag{19}$$

$\mu>0$ is an integration constant, and λ is the bifurcation parameter to be used later. We note that from (13) and (18)

$$\left(\frac{C}{\mu}\right)^{2/3} = \frac{1}{(R^-(0))^2}-\frac{\rho}{(R^+(0))^2} = c^2\left((u^-)^2+(v^-)^2-\rho((u^-)^2+(v^-)^2)\right)\big|_{\theta^\pm=0}\,. \tag{20}$$

Thus μ and λ are related to the velocity of the particle at the crest of the wave and relevant to the amplitude of the interface $y(0)$ by using (15). By (10), (11), and (18),

$$\Phi^+(\theta^+) = -\frac{1}{2\pi}\int_{-\pi}^{\pi}\log\left|\frac{\sin((\theta^++s)/2)}{\sin((\theta^+-s)/2)}\right|\left[\frac{1}{2}\frac{d}{ds}\log(V^2(s)-\rho)-\frac{I'(s)}{3I(s)}\right]ds\,. \tag{21}$$

Also from (13) and (18), we have

$$R^-(\theta^-) = \frac{(V^2(\theta^+)-\rho)^{1/2}}{V(\theta^+)I^{1/3}(\theta^+)}\,. \tag{22}$$

By (10) and (22),

$$\Phi^-(\theta^-) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\log\left|\frac{\sin((\theta^-+s^-)/2)}{\sin((\theta^--s^-)/2)}\right|\left[\frac{d}{ds^-}\log\frac{(V^2(s^+)-\rho)^{1/2}}{V(s^+)I^{1/3}(s^+)}\right]ds^-\,, \tag{23}$$

where $s^+ = L^{-1}(s^-)$. From (13) and (14), we have that $\Phi^+(\theta^+) = \Phi^-(\theta^-) = \Phi^-(L(\theta^+))$. Therefore, (21) and (23) determine the solutions $\Phi^+(\theta^+)$ and $V(\theta^+)$. In the following, it is convenient to let $\theta \stackrel{\text{def}}{=} \theta^+$ and $\Phi(\theta) \stackrel{\text{def}}{=} \Phi^+(\theta^+)$.

Since (13) and (14) imply that $L(\theta^+)$ is odd for $\theta\in[-\pi,\pi]$, by (3) we consider only the solution $\Phi(\theta)\in[0,\pi]$ for $\theta\in[-\pi,0]$. Thus $I(\theta)$ in (19) is well defined and it is straightforward to see that finding $\varphi^\pm(x,y)$ and $\psi^\pm(x,y)$ is equivalent to finding $\Phi(\theta)$ and $V(\theta)$ satisfying (21) and (23). It is noted that if $\rho=0$, then (23) with $V=1, \theta^-=\theta$ is reduced to the integral equation obtained by Nekrasov [3]. In the

following, we assume $0 < \rho < 1$. Under this condition, we can solve $V(\theta)$ in terms of $\Phi(\theta)$ from (21).

Define an integral operator $H[f]$, called the Hilbert transform (see more discussion in [26]), by

$$H[f] = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{f(s)}{\tan((\theta-s)/2)} ds\,,$$

where $f(x) \in C^{\kappa}$ with $0 < \kappa < 1$ and $C^{n,\kappa}$ is the classical Hölder space [26] with $C^{0,\kappa} = C^{\kappa}$. The transform has the following properties:

$$H[H[f]] = -f + \frac{1}{2\pi}\int_{-\pi}^{\pi} f(s)ds, \tag{24}$$

and if $f(x), f'(x)$ are continuous for $x \in [-\pi, \pi]$ and $f(x) = f(-x)$, then

$$\begin{aligned} &\frac{1}{2\pi}\int_{-\pi}^{\pi} \log\left|\frac{\sin((\theta+s)/2)}{\sin((\theta-s)/2)}\right| f_s(s)ds \\ &\qquad = -\frac{1}{\pi}\int_{-\pi}^{\pi} \log|\sin((\theta-s)/2)| f_s(s)ds = H[f]\,. \end{aligned} \tag{25}$$

Applying H on both sides of (21) and using (24), we obtain

$$H[\Phi](\theta) = \log\frac{(V^2-\rho)^{1/2}(C\mu^{-1})^{1/3}}{(1-\rho)^{1/2}I^{1/3}} - \frac{1}{2\pi}\int_{-\pi}^{\pi} \log\frac{(V^2-\rho)^{1/2}(C\mu^{-1})^{1/3}}{(1-\rho)^{1/2}I^{1/3}} ds\,. \tag{26}$$

If we let

$$\alpha = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log\frac{(1-\rho V^2)^{1/2}(C\mu^{-1})^{1/3}}{(1-\rho)^{1/2}I^{1/3}} ds\,,$$

then (26) becomes

$$W(\theta) = (1-\rho)^{1/2}(C\mu^{-1})^{-1/3} I^{1/3} \exp(\alpha + H[\Phi])\,,$$

where $W(\theta) = (V^2(\theta) - \rho)^{1/2}$. Thus by the definition of I in (19),

$$W^3(\theta) = (1-\rho)^{3/2}\mu\left(\mu^{-1} - \int_0^{\theta} \sin\Phi(s)W(s)ds\right)\exp\Big(3\alpha + 3H[\Phi](\theta)\Big)\,. \tag{27}$$

Multiplying both sides of (27) by $\exp(-3H[\Phi](\theta))$ and differentiating it with respect to θ, we get

$$(W^2(\theta))' - 2W^2(\theta)H'[\Phi](\theta) = -(2(1-\rho)^{3/2}\mu/3)\Big(\sin\Phi(\theta)\Big)\exp\Big(3\alpha + 3H[\Phi](\theta)\Big)\,,$$

which implies

$$W^2(\theta) = e^{2H[\Phi](\theta)}\left(D - (2(1-\rho)^{3/2}\mu/3)\int_0^{\theta} (\sin\Phi(s))e^{3\alpha + H[\Phi](s)}ds\right)\,,$$

where $D = (V^2(0)-\rho)^2 e^{-2H[\Phi](0)}$. From (26), we let $\theta = 0$ and obtain that $D = (1-\rho)e^{2\alpha}$. Therefore,

$$V^2(\theta) - \rho = e^{2H[\Phi](\theta)}A(\theta)\,, \tag{28}$$

where $A(\theta)$ is defined as follows:

$$A(\theta) \stackrel{\text{def}}{=} (1-\rho)e^{2\alpha} - (2(1-\rho)^{3/2}\mu/3)\int_0^\theta (\sin\Phi(s))e^{3\alpha+H[\Phi](s)}ds\,. \tag{29}$$

$V(\theta)$ and $I(\theta)$ are then solved in terms of $\Phi(\theta)$:

$$V(\theta) = \left(\rho + e^{2H[\Phi](\theta)}A(\theta)\right)^{1/2}, \tag{30}$$

$$I(\theta) = C\left(\mu^{-1} - \int_0^\theta \sin\Phi(s)e^{H[\Phi](s)}A^{1/2}(s)ds\right). \tag{31}$$

Here α is determined by the condition that $\theta^+ = -\pi$ at $\theta^- = -\pi$, i.e.,

$$\int_0^{-\pi} V(s)ds = -\pi = \int_0^{-\pi}\left(\rho + e^{2H[\Phi](\theta)}A(\theta)\right)^{1/2}d\theta\,. \tag{32}$$

There is a unique $\alpha \in (-\infty, +\infty)$ such that (32) is satisfied since the right-hand side of (30) is a strictly increasing function of α and goes to $\sqrt{\rho} < 1$ or $+\infty$ as $\alpha \to -\infty$ or $+\infty$.

(10) and (25) give

$$\Phi^-(\theta^-) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\log R^-(s^-)}{\tan((\theta^- - s^-)/2)}ds^-\,.$$

By taking the Hilbert transform on both sides of the equation, we obtain

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\Phi^-(s^-)}{\tan((\theta^- - s^-)/2)}ds^- = -\log R^-(\theta^-) + \frac{1}{2\pi}\int_{-\pi}^{\pi}\log R^-(s^-)ds^-\,,$$

which yields

$$R^-(\theta^-) = \exp\left(\gamma - \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\Phi(s)V(s)}{\tan((L(\theta)-L(s))/2)}ds\right), \tag{33}$$

where $\gamma = \frac{1}{2\pi}\int_{-\pi}^{\pi}\log R^-(s^-)ds^-$.

Next we need to rewrite the equation of $V(\theta)$ in (30). From (21), (23), and (25),

$$\Phi^+(\theta^+) = -H\left[\log R^+(\theta^+)\right], \tag{34}$$

$$\Phi^-(\theta^-) = -\frac{1}{\pi}\int_{-\pi}^{\pi}\left(\log\left|\sin((\theta^- - s^-)/2)\right|\right)\left(\frac{d}{ds^-}\log\frac{R^+(s^+)}{V(s^+)}\right)ds^-\,. \tag{35}$$

By the relations between $\theta^+, \theta^-, \Phi^+$, and Φ^-, (35) can be transformed into

$$\begin{aligned}\Phi(\theta) &= -\frac{1}{\pi}\int_{-\pi}^{\pi}\log\left|\sin\frac{L(\theta)-L(s)}{2}\right|\left(\frac{d}{ds}\log\frac{R^+(s)}{V(s)}\right)ds\\ &= H\left[\log\frac{R^+(s)}{V(s)}\right] + P[V,\Phi]\,,\end{aligned} \tag{36}$$

where by (33)

$$\begin{aligned}P[V,\Phi] &= -\frac{1}{\pi}\int_{-\pi}^{\pi}\log\left|\frac{\sin((L(\theta)-L(s))/2)}{\sin((\theta-s)/2)}\right|\left(\frac{d}{ds}\log R^-(s^-)\right)ds\\ &= \frac{1}{\pi}\int_{-\pi}^{\pi}\log\left|\frac{\sin((L(\theta)-L(s))/2)}{\sin((\theta-s)/2)}\right|\left(\frac{d}{ds}\left(\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\Phi(t)V(t)}{\tan((L(s)-L(t))/2)}dt\right)\right)ds\,.\end{aligned}$$

Adding (34) to (36), we obtain $2\Phi(\theta) = -H[\log V(\theta)] + P[V,\Phi]$ or

$$2H[\Phi](\theta) = \log V(\theta) - \frac{1}{2\pi}\int_{-\pi}^{\pi} \log V(\theta)d\theta + H[P[V,\Phi]]\,,$$

which yields $V(\theta) = \exp(2H[\Phi] - H[P[V,\Phi]] + \beta)$ with $\beta = (1/2\pi)\int_{-\pi}^{\pi}\log V(\theta)d\theta$. By the condition (32), β is obtained as follows:

$$e^{\beta} = \pi\left(\int_{-\pi}^{0} e^{2H[\Phi]-H[P[V,\Phi]]}d\theta\right)^{-1}. \tag{37}$$

Also from (28), we have $e^{4H[\Phi]-2H[P[V,\Phi]]+2\beta} - \rho - e^{2H[\Phi]}A(\theta) = 0$, which can be transformed into

$$e^{2H[\Phi]} = \frac{A(\theta) + \sqrt{A^2(\theta) + 4\rho e^{2\beta-2H[P[V,\Phi]]}}}{2e^{2\beta-2H[P[V,\Phi]]}}$$

or

$$\Phi(\theta) = -\frac{1}{2}H\left[\log\left(\frac{A(\theta) + \sqrt{A^2(\theta) + 4\rho e^{2\beta-2H[P[V,\Phi]]}}}{2e^{2\beta-2H[P[V,\Phi]]}}\right)\right] \stackrel{\text{def}}{=} \mathcal{W}[\Phi]\,, \tag{38}$$

where β is determined in (37).

Note that the function $V(\theta)$ in $P[V,\Phi]$ can be expressed in terms of $\Phi(\theta)$ using (30). On one hand, for (30) to be valid, the term inside the square root must not be negative. Also from (18), $I(\theta)$ in (31) cannot be zero. By checking $A(\theta)$ in (29) and $I(\theta)$, we have that $A(\theta)$ is always positive and $I(\theta)$ is never zero if $\Phi(\theta) \in [0,\pi]$ for $\theta \in [-\pi, 0]$, which is true in our case (see Lemma 2 below). On the other hand, in order to apply the global bifurcation theory, we need to define operators with no restriction $\Phi \in [0,\pi]$. Therefore, we need cut-off functions to modify our operator so that $V(\theta)$ in (30) is well-defined and $I(\theta)$ in (31) is not equal to zero for any $\Phi(\theta) \in C^{1,\kappa}$.

4. Modified equations using a cut-off function. We rewrite $A(\theta)$ and $I(\theta)$ in (29) and (31) as follows:

$$\tilde{A}_0(\theta) \stackrel{\text{def}}{=} e^{2\alpha}(1-\rho)\left[1 + \left(2(1-\rho)^{1/2}/3\right)E\left(\int_{\theta}^{0}\mu\sin(\Phi(s))e^{\alpha+H[\Phi](s)}\,ds\right)\right], \tag{39}$$

$$\tilde{I}_0(\theta) \stackrel{\text{def}}{=} (C/\mu)\left(1 + E\left(\int_{\theta}^{0}\mu\sin\Phi(s)e^{H[\Phi](s)}\tilde{A}_0^{1/2}(s)\,ds\right)\right), \tag{40}$$

where $E(y)$ is an infinitely differentiable function on $\mathbf{R}$ with $E(y) = y$ for $y \geq -\delta$ and $E(y) = -2\delta$ for $y \leq -2\delta$ such that $0 \leq E'(y) \leq 1$ and $|E^{(n+1)}(y)| \leq K\delta^{-n}$ for $n \geq 1$. Here $\delta > 0$ is chosen so that $4\delta = \min((3/2)(1-\rho)^{-1/2}, 1)$. Thus, $\tilde{A}_0(\theta) \geq (1/2)(1-\rho)e^{2\alpha}$ and $\tilde{I}_0(\theta) \geq (C/2\mu)$ for any Hölder continuous function $\Phi(\theta)$. Obviously, if $\pi \geq \Phi(\theta) \geq 0$, then $\tilde{A}_0(\theta) = A(\theta)$ and $\tilde{I}_0(\theta) = I(\theta)$.

By (30), we define $\tilde{V}_0(\theta) \stackrel{\text{def}}{=} (\rho + e^{2H[\Phi](\theta)}\tilde{A}_0(\theta))^{1/2}$. Then $\tilde{V}_0(\theta) > \rho^{1/2}$. Here α in (39) is determined by the condition

$$\int_{-\pi}^{0}\left(\rho + e^{2H[\Phi](\theta)}\tilde{A}_0(\theta)\right)^{1/2}d\theta = \pi\,.$$

If $y = e^{\alpha}$, $\tilde{A}_0(\theta)$ is a strictly increasing function of y since for any fixed θ and $y > 0$ by the definition of $E(y)$

$$\begin{aligned}(d\tilde{A}_0/dy) &= 2y(1-\rho)\left[1+\left(2(1-\rho)^{1/2}/3\right)E\left(\int_\theta^0 \mu\sin(\Phi(s))ye^{H[\Phi](s)}\,ds\right)\right]\\ &\quad +y^2(1-\rho)\left(2(1-\rho)^{1/2}/3\right)E'\left(\int_\theta^0 \mu\sin(\Phi(s))ye^{H[\Phi](s)}\,ds\right)\\ &\quad \times\int_\theta^0 \mu\sin(\Phi(s))e^{H[\Phi](s)}ds\\ &\geq 2y(1-\rho)(1/2)-y(1-\rho)^{3/2}(2/3)2\delta\\ &> y(1-\rho)(1-(4/3)(1-\rho)^{1/2}\delta\pi)\geq y(1-\rho)(1/2)>0\,.\end{aligned}$$

When $y\to 0$, $\int_{-\pi}^0(\rho+e^{2H[\Phi](\theta)}\tilde{A}_0(\theta))^{1/2}d\theta\to\rho^{1/2}\pi<\pi$. As $y\to+\infty$, $\int_{-\pi}^0(\rho+e^{2H[\Phi](\theta)}\tilde{A}_0(\theta))^{1/2}d\theta\to+\infty$. Thus, there must be a unique $y\in(0,\infty)$ and hence a unique $\alpha\in(-\infty,+\infty)$ with $y=e^{\alpha}$ such that

$$\int_{-\pi}^0\left(\rho+e^{2H[\Phi](\theta)}\tilde{A}_0(\theta)\right)^{1/2}d\theta=\pi\,.$$

Let this unique α be $\Lambda(\Phi)$. By using the implicit function theorem, we have that $\Lambda(\Phi)$ is Frechet differentiable with respect to Φ. Thus, if we let $\alpha=\Lambda(\Phi)$ in (39) and (40),

$$A_0(\theta)\stackrel{\text{def}}{=}e^{2\Lambda}(1-\rho)\left[1+\left(2(1-\rho)^{1/2}/3\right)E\left(\int_\theta^0\mu\sin(\Phi(s))e^{\Lambda+H[\Phi](s)}\,ds\right)\right],\tag{41}$$

$$I_0(\theta)\stackrel{\text{def}}{=}(C/\mu)\left(1+E\left(\int_\theta^0\mu\sin\Phi(s)e^{H[\Phi](s)}A_0^{1/2}(s)\,ds\right)\right),\tag{42}$$

then

$$V_0(\theta)\stackrel{\text{def}}{=}\left(\rho+e^{2H[\Phi](\theta)}A_0(\theta)\right)^{1/2}\tag{43}$$

always satisfies $\int_{-\pi}^0 V_0(\theta)d\theta=\pi$ for any $\Phi\in C^{1,\kappa}$ and $V_0(\theta)$ is a function of Φ only. We note that $(\min_{-\pi\leq\theta\leq\pi}V_0(\theta))^{-1}\leq\rho^{-1/2}$. Hence, if $\|\Phi\|_{C^{1,\kappa}}\leq K_0$, then $V_0(\theta)\in C^{1,\kappa}$ with $\|V_0(\theta)\|_{C^{1,\kappa}}\leq K$. We summarize this in the following theorem.

THEOREM 2. *For a given $\Phi(\theta)\in C^{1,\kappa}$ with $\|\Phi(\theta)\|_{C^{1,\kappa}}\leq K_0$, then $V_0(\theta)\in C^{1,\kappa}$ with $\|V_0(\theta)\|_{C^{1,\kappa}}\leq K$ and $\min_{-\pi\leq\theta\leq\pi}V_0(\theta)\geq\rho^{1/2}$, where K is a constant that is independent of Φ but may depend on K_0. If $0\leq\Phi\leq\pi$, then $V_0(\theta)\equiv V(\theta)$.*

Finally, using (41)–(43), we rewrite $P[\Phi]$ in (36) by

$$\begin{aligned}P_0[\Phi]=\frac{1}{2\pi^2}\int_{-\pi}^{\pi}\log\left|\frac{\sin((L_0(\theta)-L_0(s))/2)}{\sin((\theta-s)/2)}\right|\\ \times\frac{d}{ds}\left(\int_{-\pi}^{\pi}\frac{\Phi(t)V_0(t)dt}{\tan((L_0(s)-L_0(t))/2)}\right)ds\,,\end{aligned}\tag{44}$$

where $L_0(\theta)=\int_0^\theta V_0(s)ds$ and $V_0(\theta)$ in (43) depends only upon Φ. Then (38) is changed to

$$\Phi(\theta)=-\frac{1}{2}H\left[\log\left(\frac{A_0(\theta)+\sqrt{A_0^2(\theta)+4\rho e^{2\beta-2H[P_0]}}}{2e^{2\beta-2H[P_0]}}\right)\right]\stackrel{\text{def}}{=}\mathcal{W}_0[\Phi]\,,\tag{45}$$

where β is defined by

$$e^{\beta} = \pi \left(\int_{-\pi}^{0} e^{2H[\Phi]-H[P_0]} d\theta \right)^{-1}.$$

If the condition for $\Phi(\theta)$ in Theorem 2 is satisfied, then $\mathcal{W}_0[\Phi](\theta) = \mathcal{W}[\Phi](\theta)$, as defined in (38).

5. Existence proof. First we try to find the corresponding linear operator for $\mathcal{W}_0$ for small Φ. From (41), (42), and (44), (45) can then be rewritten as

$$\begin{aligned} \Phi(\theta) &= H\left[(1-\rho)^{3/2}\mu(3(1+\rho))^{-1}\int_0^{\theta}\Phi(s)ds\right] + N_1[\mu,\Phi] \\ &= \lambda\left(\frac{1}{2\pi}\int_{-\pi}^{\pi}\log\left|\frac{\sin((\theta+s)/2)}{\sin((\theta-s)/2)}\right|\Phi(s)ds + \mathcal{N}[\lambda,\Phi]\right) \\ &= \lambda\left(\mathcal{L}[\Phi] + \mathcal{N}[\lambda,\Phi]\right) = \lambda\mathcal{T}[\lambda,\Phi] \end{aligned} \tag{46}$$

where λ is defined as in (19), $\mathcal{N}[\lambda,\Phi]$ is nonlinear in Φ with $\mathcal{N}(\lambda,0)=0$ and

$$\lim_{\|\Phi\|_{C^{1,\kappa}}\to 0}\left(\frac{\|\mathcal{N}[\lambda,\Phi]\|_{C^{1,\kappa}}}{\|\Phi\|_{C^{1,\kappa}}}\right) = 0.$$

Here we note that $\mathcal{T}$ in (46) may not be a positive operator. Define Banach spaces as follows.

$$\begin{aligned} B_e^{n,\kappa} = \{&f(\theta)\in C^{n,\kappa}[-\pi,\pi] \mid f(\theta)=f(-\theta) \text{ for } \theta\in[-\pi,\pi] \text{ and periodic in } \theta \\ &\text{with period } 2\pi \text{ with } f'(0)=f'(-\pi)=0\,,\ \ \|f\|_{B_e^{n,\kappa}} = \|f\|_{C^{n,\kappa}[-\pi,\pi]} < +\infty\}\,, \\ B_o^{n,\kappa} = \{&f(\theta)\in C^{n,\kappa}[-\pi,\pi] \mid f(\theta)=-f(-\theta) \text{ for } \theta\in[-\pi,\pi] \text{ and periodic in } \theta \\ &\text{with period } 2\pi \text{ with } f(0)=f(-\pi)=0\,,\ \ \|f\|_{B_o^{n,\kappa}} = \|f\|_{C^{n,\kappa}[-\pi,\pi]} < +\infty\}\,, \end{aligned}$$

where $0<\kappa<1$. For $n=1$, we denote $B_e^{n,\kappa}=B_e$ and $B_o^{n,\kappa}=B_o$.

The fact that $\mathcal{T}$ is compact in B_o can be obtained in the following lemma, whose proof will be given in Appendix B.

LEMMA 1. *$\mathcal{T}$ and $\mathcal{N}$ in* (46) *are compact operators in the Banach space B_o. Moreover, for any Φ with $\|\Phi\|_{C^{1,\kappa}} \le K_0$, then $\|\mathcal{T}[\lambda,\Phi]\|_{C^{1,\kappa_1}} + \|\mathcal{N}[\lambda,\Phi]\|_{C^{1,\kappa_1}} \le K$, where $\kappa<\kappa_1\le 1$ and K is a constant only dependent of K_0 and λ with $|\lambda|<+\infty$.*

The integral operator $\mathcal{L}[\Phi]$ has been studied by Nekrasov [3] and Keady and Norbury [7]. It can be easily shown that $\mathcal{L}[\Phi]$ has eigenvalue $\lambda_n = n$ and the corresponding eigenfunction $e_n(\theta) = -\sin(n\theta)$ for $n=1,2,3,\ldots$, and the multiplicity of each eigenvalue is one.

For the nonlinear equation (46), we have small amplitude solutions by applying a local bifurcation theory [22].

THEOREM 3. *There is a constant $h>0$ such that* (46) *has exactly one solution $\Phi(\theta)$ with $\|\Phi(\theta)\|_{C^{1,\kappa}} \le h$ for each $\lambda>1$ but near 1 and $\Phi(\theta)>0$ for $\theta\in(-\pi,0)$, and the branch of solutions (λ,Φ) bifurcates from $(1,0)$. Also, the solution can be expanded into a power series in terms of $\lambda-1$.*

Remark. In fact, each $\lambda_n = n$ is a bifurcation point and for $\lambda-\lambda_n$ small, the solution Φ on the bifurcation branch can be written as $\Phi(\theta) = \epsilon e_n(\theta) + O(\epsilon^2), \Phi'(\theta) = \epsilon e_n'(\theta) + O(\epsilon^2)$, where $\epsilon = O(\lambda-\lambda_n)$. Therefore, for λ sufficiently close to λ_n with

$n \geq 2$, the solutions on the bifurcation branch change sign if $\theta \in [-\pi, 0]$, while for λ sufficiently close to $\lambda_0 = 1$, the solutions are in $(0, \pi)$ if $\theta \in (-\pi, 0)$.

We note that the corresponding result for the solutions of a one-layered fluid was obtained by Nekrasov [3] using integral equations and Levi-Civita [1] using differential equations. Also by the definition of λ in (see [19]) and the relation (see [20]), the wave speed c must be near some critical value in order to have such a bifurcation.

To study the global bifurcation for the solution of (46), we need to use the following global bifurcation theory [21, 22]. Consider the global solution behavior of the bifurcating branches of the equation

$$X = \lambda(\mathcal{L}[X] + \mathcal{N}[X]) \qquad \text{for} \quad \lambda \in \mathbf{R} \quad \text{and} \quad X \in \mathbf{B}\,,$$

where $\mathbf{B}$ is a Banach space over $\mathbf{R}$.

THEOREM 4. *Assume that the operators $\mathcal{L}$ and $\mathcal{N}$ from $\mathbf{B}$ to $\mathbf{B}$ are compact where $\mathcal{L}$ is linear and $\|\mathcal{N}[X]\|/\|X\| \to 0$ as $X \to 0$, and a real number λ_0^{-1} is an eigenvalue of $\mathcal{L}$ with odd algebraic multiplicity. Let $\mathcal{C}(\lambda_0)$ be a solution component in $\mathbf{R} \times \mathbf{B}$ which begins at $(\lambda, X) = (\lambda_0, 0)$. Then there are exactly two possibilities:* (i) *$\mathcal{C}(\lambda_0)$ is unbounded;* (ii) *$\mathcal{C}(\lambda_0)$ is compact and also contains a further point of the trivial solution branch in addition to $(\lambda_0, 0)$.*

Now we consider (46). Let $\mathbf{B} = B_o$ be the Banach space and $\mathcal{L}$ be defined in (46). Since $\mathcal{L}$ is a compact operator in B_o [3, 7], by Lemma 1, $\mathcal{N}$ in (46) is a compact operator in B_o. Therefore, we can apply Theorem 4 to (46) and obtain a global bifurcation branch for each eigenvalue of $\mathcal{L}$. Let us denote the bifurcation branch for the first eigenvalue by $\mathcal{C}(1)$. It is obvious that $\mathcal{C}(1)$ starts from the solutions obtained in Theorem 3 and is connected and closed. By Theorem 4, $\mathcal{C}(1)$ is either unbounded in $\mathbf{R} \times C^{1,\kappa}$ or compact with $(1/n, 0)$ in $\mathcal{C}(1)$ for some $n \geq 2$.

LEMMA 2. *If $(\lambda, \Phi) \in \mathcal{C}(1)$, then $0 < \Phi(\theta) < \pi$ for $\theta \in (-\pi, 0)$.*

Proof. Let $\mathcal{S}_1 = \{(\lambda, \Phi) \in \mathcal{C}(1) \mid 0 \leq \Phi(\theta) \leq \pi \text{ for all } \theta \in [-\pi, 0]\}$ and $\mathcal{S}_2 = \{(\lambda, \Phi) \in \mathcal{C}(1) \mid \Phi(\theta) \notin [0, \pi] \text{ for some } \theta \in [-\pi, 0]\}$. Obviously, $\mathcal{C}(1) = \mathcal{S}_1 \cup \mathcal{S}_2$, $\mathcal{S}_1$ is closed, and $\mathcal{S}_1, \mathcal{S}_2$ are disjoint. By Theorem 1, $\mathcal{S}_1 = \{(\lambda, \Phi) \in \mathcal{C}(1) \mid 0 < \Phi(\theta) < \pi \text{ for all } \theta \in (-\pi, 0) \text{ and } \Phi'(-\pi) > 0, \Phi'(0) < 0\}$, which is an open set in $\mathcal{C}(1)$. By Theorem 3, $\mathcal{S}_1$ is not empty. Therefore, $\mathcal{C}(1) = \mathcal{S}_1$ since $\mathcal{C}(1)$ is connected. This proves the lemma. □

By Lemma 2 and the remark after Theorem 3, $\mathcal{C}(1)$ cannot include $(1/n, 0)$ for any $n \geq 2$ since the solutions near such point will change sign for $-\pi \leq \theta \leq 0$. Therefore, $\mathcal{C}(1)$ is unbounded in $\mathbf{R} \times C^{1,\kappa}$ and we have the existence of large amplitude solutions (λ, Φ) of (46) until either $\lambda \to +\infty$ or $\|\Phi\|_{C^{1,\kappa}} \to +\infty$ or both. By Lemma 2, the solutions on $\mathcal{C}(1)$ satisfy that $0 < \Phi(\theta) < \pi$ for $-\pi < \theta < 0$. Hence, solutions on $\mathcal{C}(1)$ also satisfy (38), which implies that there is an unbounded branch of solutions in $\mathbf{R} \times C^{1,\kappa}$ for (38) bifurcating from the steady state $(\lambda, \Phi(\theta)) = (1, 0)$.

After $\Phi(\theta) = \Phi^+(\theta^+)$ is determined for the corresponding λ in (38), the function $V(\theta^+)$ can be found in (30) with α determined by (32) and $\theta^- = \int_0^{\theta^+} V(s)ds = L(\theta^+)$ transforms θ^+ to θ^-. Hence, at $\theta^- = L(\theta^+)$, $\Phi^-(\theta^-) = \Phi^+(\theta^+)$ is obtained. Finally, $R^-(\theta^-)$ and $R^+(\theta^+)$ are determined by (18) and (22) and the solutions for (13), (14), and (16) are obtained. We summarize the result as follows.

THEOREM 5. *There exists a branch of solutions $(\lambda, \Phi^\pm(\theta^\pm), R^\pm(\theta^\pm))$ satisfying* (3), (13), (15), *and* (16) *with $\lambda \geq 1$ and $0 < \Phi^-(\theta^-) < \pi$ for $-\pi < \theta^- < 0$, where $\lambda = (1-\rho)^{3/2}\mu(3(1+\rho))^{-1}$, μ satisfies*

$$\frac{1}{2}\left(\frac{3g(1-\rho)\tau}{2\pi c^2 \mu}\right)^{2/3} + \frac{g(1-\rho)}{c^2} y(0) = \frac{1-\rho}{2}\,,$$

and $y(0)$ is the height of the interface at $x = 0$. The branch bifurcates from steady state $(\lambda,\ \Phi^{\pm}(\theta^{\pm}),\ R^{\pm}(\theta^{\pm})) = (1, 0, 0)$ when c is near its critical value. The solution branch exists until $|\lambda| + \|\Phi^{+}(\theta^{+})\|_{C^{1,\kappa}}$ becomes infinity.

It should be noted that two limiting cases, $|\lambda| \to +\infty$ and $\|\Phi^{+}(\theta^{+})\|_{C^{1,\kappa}} \to +\infty$, can happen separately (see numerical computations by Turner and VandenBroeck [15] and Grimshaw and Pullin [16]). For a one-layered fluid (i.e., $\rho^{+} = 0$), these limiting cases occur simultaneously, i.e., as $|\lambda| \to +\infty$, $\|\Phi^{-}(\theta^{-})\|_{C^{1,\theta}} \to +\infty$. Also we have that $\Phi^{-}(\theta^{-})$ is always finite in $C^{1,\kappa}$ as long as $\Phi^{+}(\theta^{+}) \in C^{1,\kappa}$, which implies that the break-up of $\Phi^{-}(\theta^{-})$ in $C^{1,\kappa}$ cannot happen before the break-up of $\Phi^{+}(\theta^{+})$. The numerical evidence shows that they break up at the same time [16].

Appendix A. First let us state two lemmas [27].

LEMMA A.1 (the Hopf lemma). *Let $S \subseteq \mathbf{R}^2$ be a domain in which a nonconstant function $\Phi(x, y) \in C^1(\bar{S})$ satisfies $\Delta\Phi = 0, \Phi \geq 0$. If $(\xi, \eta) \in \partial S$ is a boundary point at which the boundary has a well-defined tangent line, and if $\Phi(\xi, \eta) = 0$, then $\partial_n \Phi(\xi, \eta) < 0$ for n, the exterior normal vector at (ξ, η) to ∂S.*

LEMMA A.2 (the Hopf "corner-point" lemma). *Consider $S \subseteq \mathbf{R}^2$ and $\Phi(x, y) \in C^1(\bar{S})$ as above, and let $(\xi, \eta) \in \partial S$ be a boundary point at which the boundary curve is not smooth but consists of two C^1 arcs meeting with an interior angle $\alpha \geq \pi/2$. (That is, interior tangent vectors t_1, t_2 to the two arcs have $t_1 \cdot t_2 \leq 0$). Let $\nu = at_1 + bt_2$ for some $a, b > 0$; the vector ν points into S. If $\Phi(\xi, \eta) = 0$, then either $\partial_\nu \Phi(\xi, \eta) > 0$ or else $\partial_\nu \Phi(\xi, \eta) = 0$ for all such ν, $t_1 \cdot t_2 = 0$, and $\partial_\nu^2 \Phi(\xi, \eta) > 0$.*

Now we can use these two lemmas to prove Theorem 1. Assume that $u = \xi + i\eta = re^{i\theta}$. Since $f^{\pm}(u) = R^{\pm}(r, \theta)e^{i\Phi^{\pm}(r,\theta)}$ are analytic and $R^{\pm} \neq 0$ for $|u| < 1$ and $|u| > 1$, $\log R^{\pm}(\xi, \eta) + i\Phi^{\pm}(\xi, \eta)$ is analytic in $\eta < 0$ except at $\xi^2 + \eta^2 = 1$ and the Cauchy–Riemann equations hold for $\log R^{\pm}$ and $\Phi^{\pm}$. Let $D^{+} = \{(\xi, \eta) = re^{i\theta} : r < 1 \text{ and } \eta < 0\}$ and $D^{-} = \{(\xi, \eta) = re^{i\theta} : r > 1 \text{ and } \eta < 0\}$. On the boundary ∂D^{+} of D^{+}, we know that $0 \leq \Phi^{+}(\theta^{+}) \leq \pi$ and $\Phi^{+}(\theta^{+}) \not\equiv 0$ for $\theta^{+} \in [-\pi, 0]$, which implies $\Phi^{+} \geq 0$ in D^{+}. If there is a $\theta_0^{+} \in (-\pi, 0)$ such that $\Phi(\theta_0^{+}) = 0$, then Φ^{+} has a minimum at θ_0^{+} with $(d\Phi^{+}(\theta_0^{+})/d\theta) = 0$. By Lemma A.1, $\Phi_r^{+}(\theta_0^{+}) < 0$. Similarly, by the condition that $\Phi^{+}(\theta^{+}) = \Phi^{-}(\theta^{-})$ on $r = 1$ and $\Phi^{-} \to 0$ as $r \to +\infty$, we have $\Phi_r^{-}(\theta_0^{-}) > 0$, where θ_0^{+} and θ_0^{-} correspond to a single point on $r = 1$. By using $r^2 = \xi^2 + \eta^2$, $\theta = \arctan(\eta/\xi)$, and the Cauchy–Riemann equations for $\log R^{\pm}$ and $\Phi^{\pm}$, we obtain that

$$(\log R)_\theta = -\Phi_\eta r \sin\theta - \Phi_\xi r \cos\theta\,,$$

$$\Phi_\xi = (-1/r)(\Phi_\theta \sin\theta - \Phi_r r \cos\theta)\,, \quad \Phi_\eta = (1/r)(\Phi_r r \sin\theta + \Phi_\theta \cos\theta)\,,$$

which implies $(\log R)_\theta = -r\Phi_r$. Thus, at $r = 1, \theta = \theta_0^{\pm}$, we have

$$\frac{R_{\theta^+}^{+}(\theta_0^{+})}{R^{+}(\theta_0^{+})} = -\Phi_r^{+}(\theta_0^{+}) > 0\,, \qquad \frac{R_{\theta^-}^{-}(\theta_0^{-})}{R^{-}(\theta_0^{-})} = -\Phi_r^{-}(\theta_0^{-}) < 0\,.$$

Then by (16), we obtain that the left-hand side of (16) is strictly positive, which contradicts (16). Therefore, there is no $\theta_0^{\pm}$ with $0 < \theta_0^{\pm} < \pi$ such that $\Phi^{\pm}(\theta_0^{\pm}) = 0$. By a similar proof, we can show that there is no $\theta_0^{\pm}$ with $0 < \theta_0^{\pm} < \pi$ such that $\Phi^{\pm}(\theta_0^{\pm}) = \pi$. Hence, $\Phi^{\pm}(\theta^{\pm})$ cannot reach its minimum 0 or maximum π at $0 < \theta^{\pm} < \pi$.

At $\theta^{\pm} = 0$, $\Phi^{\pm}(0) = 0$ and $\Phi^{\pm}$ take their minima in $D^{\pm}$ at $\theta^{\pm} = 0$. Thus $\Phi_{\theta^{\pm}}^{\pm}(0-) \leq 0$. Asusume that $\Phi_{\theta^{\pm}}^{\pm}(0-) = 0$. First let us consider $\Phi^{+}(r, \theta^{+})$ in D^{+}.

Obviously, $\Phi^+(r,\theta^+)$ is harmonic in D^+ with $\Phi^+_r(1,0)=\Phi^+_{\theta^+}(1,0)=0$. The point $(\xi,\eta)=(1,0)$ is a corner of D^+ and two interior tangent vectors $t_1=(0,-1), t_2=(-1,0)$ at this point have the properties that $t_1\cdot t_2=0$ and $\partial_\nu\Phi^+(\xi,\eta)|_{1,0}=0$ for any $\nu=at_1+bt_2$ with $a,b>0$. Therefore, by Lemma A.2, we have $\partial^2_\nu\Phi^+(\xi,\eta)|_{(1,0)}>0$. If we let $\nu=(-1,-1)$, then $\partial^2_\nu\Phi^+(\xi,\eta)|_{(1,0)}=\Phi^+_{\xi\eta}(1,0)>0$. At $\theta^+=0$, $\Phi^+_{\xi\eta}(1,0)=\Phi^+_{r\theta^+}(1,0)>0$. Similarly, we have that $\Phi^-_{r\theta^-}(1,0)<0$. But we know that $R^\pm_{\theta^\pm}=-rR^\pm\Phi^\pm_r$, which implies that $R^\pm_{\theta^\pm}(\xi,\eta)|_{(1,0)}=0$. From (16),

$$\frac{V(\theta^+)\Phi^-_r(1,\theta^-)}{(R^-(\theta^-))^2}-\frac{\rho\Phi^+_r(1,\theta^+)}{(R^+(\theta^+))^2}+\frac{\tau g(1-\rho)}{2\pi c^2}R^+(\theta^+)\sin\Phi^+(\theta^+)=0. \tag{A.1}$$

By taking the derivative of both sides of (A.1) with respect to θ^+ and using the fact that $R^\pm_{\theta^\pm}=\Phi^\pm_r=\Phi^\pm_{\theta^\pm}=\Phi^\pm=0$ at $(\xi,\eta)=(1,0)$, we obtain that at $(\xi,\eta)=(1,0)$,

$$\frac{V^2(\theta^+)\Phi^-_{r\theta^-}(\theta^-)}{(R^-(\theta^-))^2}-\frac{\rho\Phi^+_{r\theta^+}(\theta^+)}{(R^+(\theta^+))^2}=0\,,$$

which contradicts the inequalities $\Phi^+_{r\theta^+}>0$ and $\Phi^-_{r\theta^-}<0$ at $(\xi,\eta)=(1,0)$. Thus, $\Phi^\pm_\theta(0)<0$. Similarly, we can show that $\Phi^\pm_\theta(-\pi)>0$. The proof is complete.

Appendix B. Let a set $S\subset B_o$ such that $\Phi\in S$ if $\|\Phi\|_{B_o}\le K_0<+\infty$. It is obvious that $I_0(\theta)\in B^{1,\kappa_1}_e, A_0(\theta)\in B^{1,\kappa_1}_e$ for $1\ge\kappa_1>\kappa$ with bounded norms and $\mathcal{W}_0[\Phi]$ is odd and periodic with period 2π if $\Phi\in S$. Therefore, we need only show $P_0[\Phi]\in B^{1,\kappa_1}_e$ for any $1\ge\kappa_1>\kappa$. Consider integral

$$H_1[\Phi,V_0](\theta)=H_1(\theta)\stackrel{\text{def}}{=}\frac{1}{2\pi}\int_0^\pi\frac{\Phi(s)V_0(s)}{\tan((L_0(\theta)-L_0(s))/2)}ds\,. \tag{B.1}$$

If we let $t=u(s)=\int_0^s V_0(s)ds$ and $w=u(\theta)=\int_0^\theta V_0(s)ds$, then

$$H_1[\Phi,V_0]=\frac{1}{2\pi}\int_{-\pi}^\pi\frac{\Phi(u^{-1}(t))}{\tan((w-t)/2)}dt\stackrel{\text{def}}{=}W(w)\,.$$

The Hölder norm of W with respect to w is

$$\frac{|W(w_1)-W(w_2)|}{|w_1-w_2|^\kappa}\le K\|\Phi(u^{-1}(w))\|_{C^{0,\kappa}}\,,$$

which yields

$$\begin{aligned}\frac{|W(u(\theta_1))-W(u(\theta_2))|}{|\theta_1-\theta_2|^\kappa}&=\frac{|W(u(\theta_1))-W(u(\theta_2))|}{|u(\theta_1)-u(\theta_2)|^\kappa}\times\frac{|u(\theta_1)-u(\theta_2)|^\kappa}{|\theta_1-\theta_2|^\kappa}\\&\le K\left\|\Phi(u^{-1}(s))\right\|_{C^{0,\kappa}}\left|(\theta_1-\theta_2)^{-1}\int_{\theta_1}^{\theta_2}V_0(s)ds\right|^\kappa\le K\|\Phi(u^{-1}(s))\|_{C^{0,\kappa}}\|V_0\|^\kappa_{C^0}\,.\end{aligned}$$

The Hölder norm of $\Phi(u^{-1}(s))$ with respect to s is

$$\begin{aligned}\frac{|\Phi(u^{-1}(s_1))-\Phi(u^{-1}(s_2))|}{|s_1-s_2|^\kappa}&=\frac{|\Phi(u^{-1}(s_1))-\Phi(u^{-1}(s_2))|}{|u^{-1}(s_1)-u^{-1}(s_2)|^\kappa}\times\frac{|u^{-1}(s_1)-u^{-1}(s_2)|^\kappa}{|s_1-s_2|^\kappa}\\&\le K\|\Phi(\theta)\|_{C^{0,\kappa}}\,,\end{aligned}$$

where Theorem 2 has be used. Therefore, $\|H_1[\Phi, V_\epsilon]\|_{C^{0,\kappa}} \le K\|V_0\|^\kappa_{C^0}\|\Phi\|_{C^{0,\kappa}}$. The derivative of W with respect to w is

$$W'(w) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{\Phi'(u^{-1}(t))}{V_0(u^{-1}(t))\tan((w-t)/2)}dt\,,$$

whose Hölder norm can be estimated as

$$\frac{|W'(w_1) - W'(w_2)|}{|w_1 - w_2|^\kappa} \le K\left\|\Phi'\left(u^{-1}(t)\right)\left(V_0(u^{-1}(t))\right)^{-1}\right\|_{C^{0,\kappa}}.$$

Hence,

$$\|H_1[\Phi, V_0]\|_{C^{1,\kappa}}$$
$$\le K\left(\|V_0\|^\kappa_{C^{0,\kappa}} + \left\|\Phi'(u^{-1}(t))\left(V_0(u^{-1}(t))\right)^{-1}\right\|_{C^{0,\kappa}}\|V_0\|^\kappa_{C^0}\right).$$

But $|\Phi'(u^{-1}(t))(V_0(u^{-1}(t)))^{-1}| \le K\|\Phi'\|_{C^{0,\kappa}}$ and

$$\left|\Phi'(u^{-1}(t_1))\left(V_0(u^{-1}(t_1))\right)^{-1} - \Phi'(u^{-1}(t_2))\left(V_0(u^{-1}(t_2))\right)^{-1}\right||t_1 - t_2|^{-\kappa}$$
$$\le K\left\|\Phi'(\theta)\left(V_0(\theta)\right)^{-1}\right\|_{C^{0,\kappa}} \le K\left(\|V_0\|_{C^0}\|\Phi'\|_{C^{0,\kappa}} + \|\Phi'\|_{C^0}\|V_0\|_{C^{0,\kappa}}\right),$$

which gives

$$\|H_1[\Phi, V_0]\|_{C^{1,\kappa}} \le K\left(\|V_0\|_{C^0}\|\Phi\|_{C^{1,\kappa}} + \|\Phi\|_{C^{1,0}}\|V_0\|_{C^\kappa}\right).$$

Using Theorem 2, we can obtain

$$\|H_1[\Phi, V_0]\|_{C^{1,\kappa}} \le K\|\Phi\|_{C^{1,\kappa}}. \tag{B.2}$$

Next, we study

$$P_0[\Phi] = -\frac{1}{\pi}\int_{-\pi}^{\pi}\log\left|\frac{\sin((L_0(\theta) - L_0(s))/2)}{\sin((\theta - s)/2)}\right|\left(\frac{d}{ds}H_1(s)\right)ds\,. \tag{B.3}$$

First, we let $-\pi/2 \le \theta \le \pi/2$. Since $L_0(-\pi) = -\pi$ and $L_0(\pi) = \pi$, we have $-3\pi/4 \le (\theta - s)/2 \le 3\pi/4$ if $s \in [-\pi, \pi]$. Then there is a $\delta_0 > 0$ such that $-\pi + \delta_0 \le (L_0(\theta) - L_0(s))/2 \le \pi - \delta_0$. We rewrite (B.3) by

$$\int_{-\pi}^{\pi}\log\left|\frac{L_0(\theta) - L_0(s)}{\theta - s}\right|\left(\frac{d}{ds}H_1(s)\right)ds$$
$$+\int_{-\pi}^{\pi}\log\left|\frac{\sin((L_0(\theta) - L_0(s))/2)}{L_0(\theta) - L_0(s)}\frac{(\theta - s)}{\sin((\theta - s)/2)}\right|\left(\frac{d}{ds}H_1(s)\right)ds \stackrel{\text{def}}{=} I_1(\theta) + I_2(\theta)\,.$$

The estimates for $I_2(\theta)$ are obtained as follows:

$$\left|\frac{d\,I_2(\theta)}{d\theta}\right| \le \int_{-\pi}^{\pi}\left(\frac{d}{d\theta}\left(\log\left|\frac{\sin((L_0(\theta) - L_0(s))/2)}{L_0(\theta) - L_0(s)}\frac{\theta - s}{\sin((\theta - s)/2)}\right|\right)\right)\left(\frac{d}{ds}H_1(s)\right)ds$$
$$\le K\big(1 + \|V_0(\theta)\|_{C^0}\big)\|(dH_1(s)/ds)\|_{C^0}$$

$$\left|\frac{d^2\,I_2(\theta)}{d\theta^2}\right| \le K\big(1 + \|V_0(\theta)\|_{C^{1,0}}\big)\|(dH_1(s)/ds)\|_{C^0}.$$

The derivative of $I_1(\theta)$ can be rewritten as

$$\begin{aligned}\frac{d\,I_1(\theta)}{d\theta} &= \int_{-\pi}^{\pi} \left(\frac{V_0(\theta)}{L_0(\theta)-L_0(s)} - \frac{1}{\theta-s}\right)\left(\frac{d}{ds}H_1(s)\right) ds \\ &= \int_{-\pi}^{\pi} \frac{\int_0^1 (V_0(t\theta+(1-t)s)-V_0(\theta))\,dt}{(\theta-s)\int_0^1 V_0(t\theta+(1-t)s)dt}\left(\frac{d}{ds}H_1(s)\right) ds\,.\end{aligned}$$

Note that $|V_0(t\theta+(1-t)s)-V_0(\theta)| \le K\|V_0'\|_{C^0}|\theta-s|(1-t)$ and $|\int_0^1 V_0(t\theta+(1-t)s)ds|^{-1} \le K$, which implies

$$\left|\frac{d\,I_1(\theta)}{d\theta}\right| \le K\|V_0'\|_{C^0}\|H_1'(s)\|_{C^0}\,.$$

The estimate of the second-order derivative of $I_1(\theta)$ is

$$\begin{aligned}\left|\frac{d^2\,I_1(\theta)}{d\theta^2}\right| \le \Bigg| &\int_{-\pi}^{\pi} \frac{1}{\theta-s}\left(\frac{d}{d\theta}\frac{\int_0^1 (V_0(t\theta+(1-t)s)-V_0(\theta))\,dt}{\int_0^1 V_0(t\theta+(1-t)s)dt}\right)\left(\frac{d}{ds}H_1(s)\right) ds \\ &- \int_{-\pi}^{\pi} \frac{1}{(\theta-s)^2}\left(\frac{\int_0^1 (V_0(t\theta+(1-t)s)-V_0(\theta))\,dt}{\int_0^1 V_0(t\theta+(1-t)s)dt}\right)\left(\frac{d}{ds}H_1(s)\right) ds\Bigg| \\ = &\left|\int_{-\pi}^{\pi} \frac{g_1(\theta,s)}{\theta-s}ds\right|,\end{aligned}$$

where

$$\begin{aligned}g_1(\theta,s) \stackrel{\text{def}}{=} &\left[\left(\frac{d}{d\theta}\frac{\int_0^1 (V_0(t\theta+(1-t)s)-V_0(\theta))\,dt}{\int_0^1 V_0(t\theta+(1-t)s)dt}\right)\right. \\ &\left.-\left(\frac{\int_0^1\int_0^1 (1-t)V_0'((1-\omega)(t\theta+(1-t)s)+\theta\omega)d\omega dt}{\int_0^1 V_0(t\theta+(1-t)s)dt}\right)\right]\left(\frac{d}{ds}H_1(s)\right).\end{aligned}$$

Since $V_0(\theta)$ and $H_1(\theta)$ are in $C^{1,\kappa}$, $g_1(\theta,s)$ is in C^κ with respect to s and θ for $-\pi/2 \le \theta \le \pi/2$ and $-\pi \le s \le \pi$. If $v = \theta - s$, then $\theta+\pi \ge \pi/2$ and $\theta-\pi \le -\pi/2$ and

$$\begin{aligned}\int_{-\pi}^{\pi} \frac{g_1(\theta,s)}{\theta-s}ds &= \int_{\theta-\pi}^{\theta+\pi} \frac{g_1(\theta,\theta-v)}{v}\,dv \\ &= \left(\int_{\theta-\pi}^{-\pi/2} + \int_{-\pi/2}^{\pi/2} + \int_{\pi/2}^{\theta+\pi}\right)\frac{g_1(\theta,\theta-v)}{v}dv\,.\end{aligned}$$

By a similar argument as the proof of Lemma 3 in [28], for $-\pi/2 \le \theta \le \pi/2$, we have

$$\left\|\frac{d^2\,I_1}{d\theta^2}\right\|_{C^{0,\kappa}[-\pi/2,\pi/2]} \le K\left(1+\|V_0(\theta)\|_{C^{1,\kappa}}\right)\|dH_1(s)/ds\|_{C^{0,\kappa}}\,.$$

For $-\pi \le \theta \le -(\pi/2)$ or $\pi/2 \le \theta \le \pi$, we can use the periodic property of (B.3). If $\pi/2 \le \theta \le (3\pi/2)$, which implies $-\pi/2 \le \theta-\pi \le \pi/2$, then by $L_0(\pi) = \pi$ and

$L_0(2\pi) = 2\pi$

$$\int_{-\pi}^{\pi} \log \left| \frac{\sin((L_0(\theta) - L_0(s))/2)}{\sin((\theta - s)/2)} \right| \left(\frac{d}{ds} H_1(s) \right) ds$$
$$= \int_{-\pi}^{\pi} \log \left| \frac{\sin\left((\int_s^{\theta-\pi} V_0(t+\pi) ds)/2 \right)}{\sin((\theta - \pi - s)/2)} \right| \left(\frac{d}{ds} H_1(s+\pi) \right) ds \,.$$

Let $\theta_0 = \theta - \pi$ and $V_{0,0}(t) = V_0(t+\pi)$. We have $-\pi/2 \le \theta_0 \le \pi/2$ and $V_{0,0}(\theta) \in C^{1,\kappa}$. By the proof for the estimates of I_1 with $\theta \in [-\pi/2, \pi/2]$, we have

$$\left\| \frac{d^2 I_1}{d\theta_0^2} \right\|_{C^{0,\kappa}[-\pi/2,\pi/2]} \le K \left(1 + \|V_0(\theta)\|_{C^{1,\kappa}}\right) \left\| \frac{d}{ds} H_1(s) \right\|_{C^{0,\kappa}} .$$

Hence, we obtain

$$\left\| \frac{d^2 I_1}{d\theta^2} \right\|_{C^{\kappa}[-\pi,\pi]} \le K \left(1 + \|V_0(\theta)\|_{C^{1,\kappa}}\right) \left\| \frac{d}{ds} H_1(s) \right\|_{C^{0,\kappa}} .$$

Combining the estimates for I_1 and I_2 together, we have shown that

$$\|P_0[\Phi]\|_{C^{2,\kappa}} \le K \left(1 + \|V_0(\theta)\|_{C^{1,\kappa}}\right) \|(dH_1(s)/ds)\|_{C^{0,\kappa}} .$$

Therefore, Theorem 2, (B.1), and (B.2) yield

$$\|P_0[\Phi]\|_{C^{2,0}} \le K \left(1 + \|V_0(\theta)\|_{C^{1,\kappa}}\right) \|\Phi\|_{C^{1,\kappa}}$$
$$\le K \|\Phi\|_{C^{1,\kappa}} ,$$

where K is a constant independent of Φ, κ but may depend on K_0. Hence, $P_0[\Phi] \in B_e^{1,\kappa_1}$ for any $1 \ge \kappa_1 > \kappa$ and $\mathcal{T}$ is compact in B_o since any bounded set in C^{1,κ_1} is a compact set of $C^{1,\kappa}$ if $1 \ge \kappa_1 > \kappa$.

Acknowledgment. The author wishes to thank Professor T. B. Benjamin for introducing this problem to him.

REFERENCES

[1] T. Levi-Civita, *Détermination rigoureuse des ondes permanentes d'ampleur finie*, Math. Ann., 93 (1925), pp. 264–314.
[2] D. J. Struik, *Détermination rigoureuse des ondes irrotationnelles périodiques dans un cannal á profondeur finie*, Math. Ann., 95 (1926), pp. 595–634.
[3] A. I. Nekrasov, *The Exact Theory of Steady State Waves on the Surface of a Heavy Liquid*, MRC Technical Summary Report 813, University of Wisconsin, Madison, WI, 1967.
[4] K. O. Friedrichs and D. H. Hyers, *The existence of solitary waves*, Comm. Pure Appl. Math., 7 (1954), pp. 517–550.
[5] J. T. Beale, *The existence of solitary water waves*, Comm. Pure Appl. Math., 30 (1977), pp. 373–389.
[6] Yu. P. Krasovskii, *On the theory of steady state waves of large amplitude*, U.S.S.R. Computat. Math. Math. Phys., 1 (1961), pp. 836–855.
[7] G. Keady and J. Norbury, *On the existence theory for irrotational water waves*, Math. Proc. Cambridge Philos. Soc., 83 (1978), pp. 137–157.
[8] C. J. Amick and J. F. Toland, *On solitary water-waves of finite amplitude*, Arch. Rational Mech. Anal., 76 (1981), pp. 9–96.

[9] J. F. Toland, *On the existence of a wave of greatest height and Stokes' conjecture*, Proc. Roy. Soc. London Ser. A, 363 (1978), pp. 469–485.
[10] C. J. Amick, L. E. Fraenkel, and J. F. Toland, *On the Stokes conjecture for the wave of extreme form*, Acta Math., 148 (1982), pp. 193–214.
[11] J. B. McLeod, *The Stokes and Krasovskii conjectures for the wave of greatest height*, Stud. Appl. Math., 98 (1997), pp. 311–333.
[12] B. Buffoni, E. N. Dancer, and J. F. Toland, *The sub-harmonic bifurcation of Stokes waves*, Arch. Rational Mech. Anal., 153 (2000), pp. 241–271.
[13] J. Y. Holyer, *Large amplitude progressive interfacial waves*, J. Fluid Mech., 93 (1979), pp. 433–448.
[14] D. I. Meiron and P. G. Saffman, *Overhanging interfacial gravity waves of large amplitude*, J. Fluid Mech., 129 (1983), pp. 213–218.
[15] R. E. L. Turner and J.-M. VandenBroeck, *The limiting configuration of interfacial gravity waves*, Phys. Fluids, 29 (1986), pp. 372–375.
[16] R. H. Grimshaw and D. I. Pullin, *Extreme interfacial waves*, Phys. Fluids, 29 (1986), pp. 2802–2807.
[17] D. I. Pullin and R. H. J. Grimshaw, *Finite-amplitude solitary waves at the interface between two homogeneous fluids*, Phys. Fluids, 31 (1988), pp. 3550–3559.
[18] R. E. L. Turner and J.-M. VandenBroeck, *Broadening of interfacial solitary waves*, Phys. Fluids, 31 (1988), pp. 2486–2490.
[19] C. J. Amick and R. E. L. Turner, *A global theory of solitary waves in two-fluid systems*, Trans. Amer. Math. Soc., 298 (1986), pp. 431–481.
[20] T. B. Benjamin, *private communication*, 1993.
[21] P. H. Rabinowitz, *Some global results for nonlinear eigenvalue problems*, J. Funct. Anal., 7 (1971), pp. 487–513.
[22] E. Zeidler, *Nonlinear Functional Analysis and Its Applications* I, Springer-Verlag, New York, 1986.
[23] L. M. Milne-Thomson, *Theoretical Hydrodynamics*, Macmillan, New York, 1968.
[24] S. M. Sun, *Periodic waves in two-layer fluids of infinite depth*, in Advances in Multi-Fluid Flows, Y. Y. Renardy, A. V. Coward, D. T. Papageorgiou, and S. M. Sun, eds., SIAM, Philadelphia, 1996, pp. 339–345.
[25] W. Rudin, *Real and Complex Analysis*, McGraw-Hill, New York, 1966.
[26] A. Zygmund, *Trigonometric Series*, Cambridge University Press, London, 1968.
[27] W. Craig and P. Sternberg, *Symmetry of free-surface flows*, Arch. Rational Mech. Anal., 118 (1992), pp. 1–36,
[28] S. M. Sun, *Asymptotic behavior and symmetry of internal waves in two-layer fluids of great depth*, J. Differential Equations, 129 (1996), pp. 18–48.

SIAM J. MATH. ANAL.
Vol. 32, No. 5, pp. 1032–1049

RAISING MULTIWAVELET APPROXIMATION ORDER THROUGH LIFTING*

FRITZ KEINERT†

Abstract. Given a pair of biorthogonal, compactly supported multiwavelets, we present an algorithm for raising their approximation orders to any desired level, using one lifting step and one dual lifting step. Free parameters in the algorithm are explicitly identified, and can be used to optimize the result with respect to other criteria.

Key words. wavelets, multiwavelets, lifting, approximation order

AMS subject classification. 42C15

PII. S0036141098349509

1. Introduction.

A *refinable function vector of multiplicity* r *and dilation factor* m is a vector $\boldsymbol{\phi}^{(0)}$ of r real-valued functions

$$\boldsymbol{\phi}^{(0)}(x) = \begin{pmatrix} \phi_1^{(0)}(x) \\ \vdots \\ \phi_r^{(0)}(x) \end{pmatrix}, \qquad x \in \boldsymbol{R}, \tag{1.1}$$

which satisfies a *matrix refinement equation*

$$\boldsymbol{\phi}^{(0)}(x) = \sqrt{m} \sum_{k \in \boldsymbol{Z}} h_k^{(0)} \boldsymbol{\phi}^{(0)}(mx - k). \tag{1.2}$$

The sequence $\mathcal{H}^{(0)} = \{h_k^{(0)}\}_{k \in \boldsymbol{Z}}$ of coefficient matrices is called the *mask* of the function. We assume that only finitely many $h_k^{(0)}$ are nonzero and that all $\phi_j^{(0)}$ have compact support.

We call $\boldsymbol{\phi}^{(0)}$ a *multiscaling function* if it generates a *multiresolution approximation* (MRA) [21] of $L^2(\boldsymbol{R})$. This means that there exists a sequence of subspaces V_j, $j \in \boldsymbol{Z}$, of $L^2(\boldsymbol{R})$ with the following properties:

1. $V_j \subset V_{j+1}$,
2. $\bigcap_j V_j = \{0\}, \quad \overline{\bigcup_j V_j} = L^2(\boldsymbol{R})$,
3. $f(x) \in V_j \Leftrightarrow f(x - m^{-j}k) \in V_j, \quad k \in \boldsymbol{Z}$,
4. $f(x) \in V_j \Leftrightarrow f(mx) \in V_{j+1}$,
5. $\{\phi_j^{(0)}(x - k) : j, k \in \boldsymbol{Z}\}$ forms a Riesz basis of V_0.

In detail, property 5 means that there exist constants $0 < A \le B$ so that

$$A \sum_j \|\boldsymbol{c}_j\|_2^2 \le \Big\| \sum_j \boldsymbol{c}_j^* \boldsymbol{\phi}^{(0)}(x - j) \Big\|_2^2 \le B \sum_j \|\boldsymbol{c}_j\|_2^2 \tag{1.3}$$

for any sequence of coefficient vectors $\{\boldsymbol{c}_j\}$ with $\sum_j \|\boldsymbol{c}_j\|_2^2 < \infty$. The superscript $*$ denotes the transpose.

*Received by the editors December 21, 1998; accepted for publication (in revised form) July 24, 2000; published electronically January 19, 2001. This work was partially supported by ARC large grant 390 1404 1099 14 1 at Flinders University, Adelaide, South Australia.
http://www.siam.org/journals/sima/32-5/34950.html
†Department of Mathematics, Iowa State University, Ames, IA 50011 (keinert@iastate.edu).

Other function vectors $\boldsymbol{\phi}^{(\nu)}$, $\nu = 1, \ldots, m-1$, are called *multiwavelet functions* if $\{\phi_j^{(\nu)}(x-k) : j,k \in \boldsymbol{Z}\}$ form Riesz bases of other spaces $W_0^{(\nu)}$ so that

$$V_0 \oplus W_0^{(1)} \oplus \cdots \oplus W_0^{(m-1)} = V_1$$

and

$$\{m^{\ell/2}\phi_j^{(\nu)}(m^\ell x - k) : j,k,\ell \in \boldsymbol{Z}, \nu = 1, \ldots, m-1\}$$

forms a Riesz basis of $L^2(\boldsymbol{R})$.

These multiwavelet functions satisfy refinement equations

$$\boldsymbol{\phi}^{(\nu)}(x) = \sqrt{m} \sum_k h_k^{(\nu)} \boldsymbol{\phi}^{(0)}(mx - k). \tag{1.4}$$

We again assume that all coefficient sequences are finite and all multiwavelet functions have compact support.

In the standard literature, the word "wavelet" sometimes means an individual wavelet function, sometimes scaling function and wavelets together. To avoid ambiguity, we refer to the entire collection $\boldsymbol{\phi} = \{\boldsymbol{\phi}^{(\nu)} : \nu = 0, \ldots, m-1\}$ as a multiwavelet and to the individual $\boldsymbol{\phi}^{(\nu)}$ as multiscaling functions or multiwavelet functions.

The properties of refinable function vectors, multiscaling functions, and multiwavelet functions with dilation factor $m = 2$ are discussed in many papers. Some of the earliest occurrences are [1], [9], [11], [12]; more recent treatments include [4], [6], [14], [18], [25], [26], [28]. It is straightforward to extend these results to the case of general m, following the one-dimensional case which is discussed, for example, in [16], [27], [35].

Two multiwavelets $\boldsymbol{\phi}$, $\tilde{\boldsymbol{\phi}}$ form a biorthogonal pair if they satisfy the *biorthogonality conditions*

$$\int \phi_k^{(\mu)}(x) \tilde{\phi}_\ell^{(\nu)}(x-j)\, dx = \delta_{\mu\nu}\delta_{0j}\delta_{k\ell}, \tag{1.5}$$

where δ is the Kronecker delta.

One of the properties of a multiscaling function which has great practical interest is the *approximation order* [15], [17], [22], [23]. $\boldsymbol{\phi}^{(0)}$ has approximation order $p \geq 1$ if all powers of x up to $p-1$ can be locally written as linear combinations of its integer translates. That is, there exist vectors $\boldsymbol{y}_k^{(j)} \in \boldsymbol{R}^r$ such that for $j = 0, \ldots, p-1$

$$x^j = \sum_k \boldsymbol{y}_k^{(j)*} \boldsymbol{\phi}^{(0)}(x-k). \tag{1.6}$$

Since we assume compact support, the sum is finite for each fixed x, and there are no convergence problems.

This paper considers the following problem: Given a biorthogonal multiwavelet pair $\boldsymbol{\phi}$, $\tilde{\boldsymbol{\phi}}$ and integers $p \geq 1$, $\tilde{p} \geq 1$, find an algorithm to generate from them new multiwavelets $\boldsymbol{\phi}_{\text{new}}$, $\tilde{\boldsymbol{\phi}}_{\text{new}}$ with approximation orders p, $\tilde{p}$, respectively.

One known way to raise approximation order is through the use of two-scale similarity transforms (TSTs) [24], [25], [28], [29], [30]. Our approach uses *lifting*. As a systematic strategy for creating new multiwavelet functions, this approach dates back to [5], and in the more general context of stable multiscale representations to [2]. Under the name "lifting," these techniques were later applied in [3], [7], [8], [20], [33], [34]. Details can be found in section 3.

Compared to TSTs, the lifting approach has the following advantages:

1. lifting produces a complete new multiwavelet pair; TST produces only a new multiscaling function;
2. lifting uses no matrix division or singular matrices;
3. lifting generally produces shorter new masks than the TST algorithm.

The outline of this paper is as follows.

Sections 2, 3 introduce notation and summarize needed results from the literature. The main result can be found in Theorem 4.1 at the end of section 4. The proof is constructive, and forms the basis for a numerical algorithm. An alternative approach, based on a suggestion in [34], is presented in section 5. Implementation details for both algorithms are stated in section 6. Section 7 contains some examples.

2. Representations of multiwavelet masks. The results in this section are well known. Proofs or appropriate references can be found, e.g., in [2], [5], [6], or [25].

Throughout this paper, all calculations are based on the masks $\mathcal{H}$, $\tilde{\mathcal{H}}$ alone. In terms of masks, the biorthogonality conditions (1.5) are represented as

$$\sum_k h_k^{(\mu)} \tilde{h}_{k+mj}^{(\nu)T} = \delta_{\mu\nu}\delta_{0j} I. \tag{2.1}$$

Here and in the remainder of this paper, I denotes an identity matrix of appropriate size.

The existence of a biorthogonal pair of masks does not automatically guarantee the existence of a corresponding pair of MRAs. This raises the question of whether the new masks produced by our algorithm actually represent multiwavelets in the sense described in the introduction, or merely the coefficients of filter banks for signal processing applications. We will address this question in section 6.3.3, using the notation of [2], [5], which we briefly introduce at this point. (The multiscale representations discussed in [2] actually cover more general cases than described here.)

There are various conditions given in the literature (see, for example, [4], [10], [14], [19], [25]) which can be checked to see whether a given $\mathcal{H}^{(0)}$ gives rise to a refinable function vector $\boldsymbol{\phi}^{(0)}$ and an MRA. This corresponds to the concept of a (uniformly) *stable basis* in [2].

Given $\boldsymbol{\phi}^{(0)}$, the multiwavelet functions $\boldsymbol{\phi}^{(\nu)}$, $\nu = 1, \dots, m-1$, always exist by (1.4). They form a *stable completion* of $\boldsymbol{\phi}^{(0)}$ if

$$\{\phi_j^{(\nu)}(x-k) : \nu = 0, \dots, m-1, j, k \in \boldsymbol{Z}\}$$

forms a Riesz basis of V_1.

If

$$\{m^{\ell/2}\phi_j^{(\nu)}(m^\ell x - k) : j, k, \ell \in \boldsymbol{Z}, \nu = 1, \dots, m-1\}$$

forms a Riesz basis of $L^2(\boldsymbol{R})$, this is called *stability over all levels*.

In section 6.3.3, we will refer to these concepts as stability of $\boldsymbol{\phi}^{(0)}$, stability of $\boldsymbol{\phi}$, and stability over all levels, respectively.

The information contained in a mask $\mathcal{H}$ can be represented in various forms. We present here the two forms used in this paper.

The *symbol* of a function mask $\mathcal{H}^{(\nu)}$ is defined as

$$H^{(\nu)}(\xi) = \frac{1}{\sqrt{m}} \sum_k h_k^{(\nu)} e^{-ik\xi}, \qquad \xi \in \boldsymbol{R}. \tag{2.2}$$

In terms of symbols, the biorthogonality conditions (2.1) can be expressed as either

$$\sum_{k=0}^{m-1} H^{(\nu)}\left(\xi+\frac{2\pi}{m}k\right)\tilde{H}^{(\mu)*}\left(\xi+\frac{2\pi}{m}k\right)=\delta_{\nu\mu}I \tag{2.3}$$

or

$$\sum_{\nu=0}^{m-1} \tilde{H}^{(\nu)*}\left(\xi+\frac{2\pi}{m}k\right)H^{(\nu)}\left(\xi+\frac{2\pi}{m}\ell\right)=\delta_{k\ell}I, \tag{2.4}$$

where for complex-valued functions the superscript $*$ stands for the complex conjugate transpose.

A mask $\mathcal{H}$ satisfies *condition E* if $H^{(0)}(0)$ has a simple eigenvalue of 1, with all other eigenvalues less than 1 in modulus. Condition E is automatically satisfied if the mask generates an MRA of $L^2(\boldsymbol{R})$ with compactly supported basis functions [19], [25].

The *polyphase representation* $P(\xi)$ is the block matrix

$$P(\xi)=\begin{pmatrix} H_0^{(0)}(\xi) & H_1^{(0)}(\xi) & \dots & H_{m-1}^{(0)}(\xi)\\ H_0^{(1)}(\xi) & H_1^{(1)}(\xi) & \dots & H_{m-1}^{(1)}(\xi)\\ \vdots & \vdots & \ddots & \vdots\\ H_0^{(m-1)}(\xi) & H_1^{(m-1)}(\xi) & \dots & H_{m-1}^{(m-1)}(\xi)\end{pmatrix}, \tag{2.5}$$

where the *polyphase symbols* $H_\mu^{(\nu)}(\xi)$ are defined by

$$H_\mu^{(\nu)}(\xi)=\sum_k h_{mk+\mu}^{(\nu)}e^{-ik\xi}. \tag{2.6}$$

The normalization is chosen so that biorthogonality is equivalent to

$$P(\xi)\tilde{P}(\xi)^*=I. \tag{2.7}$$

The determinants of $P(\xi)$, $\tilde{P}(\xi)$ are trigonometric polynomials. If $\mathcal{H}$, $\tilde{\mathcal{H}}$ both have finite length, the determinants must be monomials.

If a multiwavelet has approximation order p, as defined in (1.6), then necessarily (see [15], [22], [23])

$$\boldsymbol{y}_k^{(j)}=\sum_{\ell=0}^{j}\binom{j}{\ell}k^{j-\ell}\boldsymbol{y}^{(\ell)} \tag{2.8}$$

for some vectors $\boldsymbol{y}^{(j)}\in\boldsymbol{R}^r$, $\boldsymbol{y}^{(0)}\neq\boldsymbol{0}$, and

$$\sum_{\ell=0}^{j}\binom{j}{\ell}(-i)^{j-\ell}m^\ell\boldsymbol{y}^{(\ell)*}D^{j-\ell}H^{(0)}\left(\frac{2\pi}{m}k\right)=\delta_{0k}\boldsymbol{y}^{(j)*} \tag{2.9}$$

for $j=0,\dots,p-1$ and $k=0,\dots,m-1$. D denotes the differentiation operator. We take (2.9) as the definition of approximation order for masks.

If $Y(\xi)$ is any vector of trigonometric polynomials with

$$D^jY(0)=i^{-j}\boldsymbol{y}^{(j)},\qquad j=0,\dots,p-1, \tag{2.10}$$

then another way to express (2.9) is

$$D^j \left[H^{(0)*}(\xi + \frac{2\pi}{m}k) Y(m\xi) \right] \Big|_{\xi=0} = \delta_{0k} D^j Y(0) = \delta_{0k} i^{-j} y^{(j)}. \tag{2.11}$$

The following theorem states that the approximation order of $\tilde{\mathcal{H}}$ can be determined by examining $\mathcal{H}$. It is a crucial step for the development in section 4. A partial result (the "only if" part for $\nu = 0$) was earlier derived in [32]. A similar result for multivariate wavelets can also be found in [13].

THEOREM 2.1. *Assume $\mathcal{H}$, $\tilde{\mathcal{H}}$ are biorthogonal masks. Then $\tilde{\mathcal{H}}$ has approximation order $\tilde{p}$ with vectors $\tilde{\boldsymbol{y}}^{(j)}$ if and only if*

$$\sum_{s=0}^{j} \binom{j}{s} i^{j-s} D^{j-s} H^{(\nu)}(0) \tilde{\boldsymbol{y}}^{(s)} = \delta_{0\nu} m^j \tilde{\boldsymbol{y}}^{(j)} \tag{2.12}$$

for $j = 0, \dots, \tilde{p} - 1$, $\nu = 0, \dots, m - 1$.

Proof. Assume $\tilde{Y}(\xi)$ satisfies (2.11) for the dual mask:

$$D^j \left[\tilde{H}^{(0)*}(\xi + \frac{2\pi}{m}k) \tilde{Y}(m\xi) \right] \Big|_{\xi=0} = \delta_{0k} D^j \tilde{Y}(0). \tag{2.13}$$

Take $\mu = 0$ in (2.3) and multiply by $\tilde{Y}(m\xi)$:

$$\sum_{k=0}^{m-1} H^{(\nu)} \left(\xi + \frac{2\pi}{m}k \right) \tilde{H}^{(0)*} \left(\xi + \frac{2\pi}{m}k \right) \tilde{Y}(m\xi) = \delta_{0\nu} \tilde{Y}(m\xi). \tag{2.14}$$

Differentiate j times and evaluate at $\xi = 0$:

$$\begin{aligned} \delta_{0\nu} m^j D^j \tilde{Y}(0) &= \sum_{s=0}^{j} \sum_{k=1}^{m-1} \binom{j}{s} D^{j-s} H^{(\nu)}(\frac{2\pi}{m}k) D^s \left[H^{(0)*}(\xi + \frac{2\pi}{m}k) \tilde{Y}(m\xi) \right] \Big|_{\xi=0} \\ &= \sum_{s=0}^{j} \sum_{k=1}^{m-1} \binom{j}{s} D^{j-s} H^{(\nu)}(0) D^s \tilde{Y}(0), \end{aligned} \tag{2.15}$$

which simplifies to (2.12). ∎

Remark. If the dual multiscaling function $\tilde{\boldsymbol{\phi}}^{(0)}$ has approximation order $\tilde{p}$, this implies that the multiwavelet functions have $\tilde{p}$ vanishing continuous moments, that is,

$$\int x^j \phi_k^{(\nu)}(x)\, dx = 0 \tag{2.16}$$

for $j = 0, \dots, \tilde{p} - 1$, $k = 1, \dots, r$, and $\nu = 1, \dots, m - 1$.

Equation (2.12) for $\nu = 1, \dots, m - 1$ can also be derived from the vanishing moment condition (2.16). Thus, Theorem 2.1 is a strictly algebraic version of the statement "the dual multiscaling function has approximation order $\tilde{p}$ if and only if the multiwavelet functions have $\tilde{p}$ vanishing moments."

3. Lifting. The following theorem forms the basis for the lifting procedure.

THEOREM 3.1. *If P_1, P_2 are polyphase matrices for two multiwavelets with the same multiscaling function, they are related by*

$$P_2(\xi) = \begin{pmatrix} I & 0 \\ L(\xi) & M(\xi) \end{pmatrix} P_1(\xi), \tag{3.1}$$

where L is of size $(m-1)r \times r$, M is of size $(m-1)r \times (m-1)r$. If both masks have finite length, $\det(M(\xi))$ is a monomial.

For wavelets of multiplicity 1 and dilation factor 2, this theorem dates back to [36].

As a general technique for creating stable refinable bases, the theorem was first used (in a periodic setting) in [5]. The most general version is given in [2], in the context of stable multiscale representations. A multiscale representation generalizes the concept of MRA by allowing each of the nested subspaces V_j to have its own basis Φ_j, not necessarily generated by translates and dilations from a small number of scaling functions. It is shown that any two stable completions of the same Φ_j are related in a manner similar to (3.1). (Polyphase matrices are not available in the general multiscale case, so the notation is different).

In the scalar case, Sweldens called (3.1) with $M = 1$ a *lifting step* [7], [33], [34], and showed that any wavelet can be built from the trivial polyphase matrix $P(\xi) = I$ by a finite combination of lifting steps and dual lifting steps:

$$P_{\text{new}}(\xi) = \begin{pmatrix} 1 & L(\xi) \\ 0 & 1 \end{pmatrix} P(\xi). \tag{3.2}$$

We also ignore M, since it has no effect on the scaling functions or dual scaling functions and their approximation orders. Thus, we define a multiwavelet lifting step as

$$P_{\text{new}}(\xi) = \begin{pmatrix} I & 0 \\ L(\xi) & I \end{pmatrix} P(\xi) = \begin{pmatrix} I & 0 & \dots & 0 \\ L^{(1)}(\xi) & I & & \\ \vdots & & \ddots & \\ L^{(m-1)}(\xi) & & & I \end{pmatrix} P(\xi), \tag{3.3}$$

where each $L^{(\nu)}(\xi)$ is an $r \times r$ matrix trigonometric polynomial. The effect on the dual is

$$\tilde{P}_{\text{new}}(\xi) = \begin{pmatrix} I & -L(\xi)^* \\ 0 & I \end{pmatrix} \tilde{P}(\xi). \tag{3.4}$$

In terms of the function symbols, our definition of multiwavelet lifting is equivalent to

$$\begin{aligned} &H^{(0)}_{\text{new}}(\xi) = H^{(0)}(\xi), \\ &H^{(\nu)}_{\text{new}}(\xi) = H^{(\nu)}(\xi) + L^{(\nu)}(m\xi)H^{(0)}(\xi), && \nu = 1, \dots, m-1, \\ &\tilde{H}^{(0)}_{\text{new}}(\xi) = \tilde{H}^{(0)}(\xi) - \textstyle\sum_{\nu=1}^{m-1} L^{(\nu)*}(m\xi)\tilde{H}^{(\nu)}(\xi), \\ &\tilde{H}^{(\nu)}_{\text{new}}(\xi) = \tilde{H}^{(\nu)}(\xi), && \nu = 1, \dots, m-1. \end{aligned} \tag{3.5}$$

In terms of multiscaling and multiwavelet functions, we have

$$\begin{aligned} &\boldsymbol{\phi}^{(0)}_{\text{new}}(x) = \boldsymbol{\phi}^{(0)}(x), \\ &\boldsymbol{\phi}^{(\nu)}_{\text{new}}(x) = \boldsymbol{\phi}^{(\nu)}(x) + \textstyle\sum_k L^{(\nu)}_k \boldsymbol{\phi}^{(0)}(x-k), && \nu = 1, \dots, m-1, \\ &\tilde{\boldsymbol{\phi}}^{(0)}_{\text{new}}(x) = \tilde{\boldsymbol{\phi}}^{(0)}(\xi) - \textstyle\sum_{\nu=1}^{m-1} \sum_k L^{(\nu)*}_{-k} \tilde{\boldsymbol{\phi}}^{(\nu)}(x-k), \\ &\tilde{\boldsymbol{\phi}}^{(\nu)}_{\text{new}}(x) = \tilde{\boldsymbol{\phi}}^{(\nu)}(x), && \nu = 1, \dots, m-1. \end{aligned} \tag{3.6}$$

Different but related multiwavelet lifting procedures are described in [8], [32]. Lifting for multivariate wavelets is discussed in [3] and [20].

4. Raising approximation order by lifting. In this section, we show how a single lifting step can be used to raise the approximation order of the dual multiscaling function to any desired level, while leaving the multiscaling function and its approximation order invariant.

In the scalar case, the idea of using lifting to raise the dual approximation order goes back to Sweldens' original papers [33], [34]. In the multiwavelet setting, different implementations appear in [8], [32]. Similar ideas can also be found in [13] (for multivariate wavelets) and [2] (for general multiscale approximations).

Let $\mathcal{H}$, $\tilde{\mathcal{H}}$ be a biorthogonal pair of masks, with $\mathcal{H}$ satisfying condition E.

Remark. As pointed out above, condition E is automatically satisfied for compactly supported stable $\boldsymbol{\phi}^{(0)}$, so it is a desirable property anyway. This is the reason why we impose condition E instead of the slightly weaker conditions we actually need.

Given any trigonometric matrix polynomial

$$L(\xi) = \sum_k L_k\, e^{-ik\xi}, \tag{4.1}$$

we define its *discrete moments* as

$$\Lambda_j = \sum_k k^j L_k = i^j D^j L(0), \qquad j = 0, 1, \ldots. \tag{4.2}$$

If the coefficients L_k are nonzero only for $k = k_0, \ldots, k_0 + n - 1$, and $N \geq 1$ is arbitrary, then L_k and Λ_k are related by

$$(\Lambda_0, \ldots, \Lambda_{N-1}) = (L_{k_0}, \ldots, L_{k_0+n-1})\, A, \tag{4.3}$$

where A is a block Vandermonde matrix with blocks of size $r \times r$

$$A = \begin{pmatrix} k_0^0 I & k_0^1 I & \cdots & k_0^{N-1} I \\ (k_0+1)^0 I & (k_0+1)^1 I & \cdots & (k_0+1)^{N-1} I \\ \vdots & \vdots & \ddots & \vdots \\ (k_0+n-1)^0 I & (k_0+n-1)^1 I & \cdots & (k_0+n-1)^{N-1} I \end{pmatrix}. \tag{4.4}$$

Let $M_j^{(\nu)}$ denote the jth discrete moment of $H^{(\nu)}$. It follows from differentiating (3.5) and evaluating at $\xi = 0$ that the new moments after lifting are given by

$$\begin{aligned} M_{\mathrm{new},j}^{(0)} &= M_j^{(0)}, \\ M_{\mathrm{new},j}^{(\nu)} &= M_j^{(\nu)} + \sum_{s=0}^{j} \binom{j}{s} m^s \Lambda_s^{(\nu)} M_{j-s}^{(0)}. \end{aligned} \tag{4.5}$$

We want to satisfy the dual approximation order conditions (2.12) of order $\tilde{p}$. We do this first for $\nu = 0$, where the conditions are

$$\sum_{s=0}^{j} \binom{j}{s} M_{j-s}^{(0)} \tilde{\boldsymbol{y}}^{(s)} = m^j \tilde{\boldsymbol{y}}^{(j)}, \qquad j = 0, \ldots, \tilde{p} - 1. \tag{4.6}$$

We can rewrite this in the form

$$\begin{aligned}\tilde{\boldsymbol{y}}^{(0)} &= M_0^{(0)}\tilde{\boldsymbol{y}}^{(0)},\\ \tilde{\boldsymbol{y}}^{(j)} &= \left(m^j I - M_0^{(0)}\right)^{-1}\sum_{s=0}^{j-1}\binom{j}{s}M_{j-s}^{(0)}\tilde{\boldsymbol{y}}^{(s)}, \qquad j=1,2,\ldots,\tilde{p}-1.\end{aligned} \tag{4.7}$$

Condition E is sufficient to guarantee solvability.

After the $\tilde{\boldsymbol{y}}^{(j)}$ have been determined, define

$$\boldsymbol{z}_j^{(\nu)} = m^{-j}\sum_{s=0}^{j}\binom{j}{s}M_{j-s}^{(\nu)}\tilde{\boldsymbol{y}}^{(s)}. \tag{4.8}$$

By (4.6),

$$\boldsymbol{z}_j^{(0)} = \tilde{\boldsymbol{y}}^{(j)} \tag{4.9}$$

for $j=0,\ldots,\tilde{p}-1$. By (2.12), $\tilde{\mathcal{H}}$ has existing approximation order $\tilde{q}$ if and only if

$$\boldsymbol{z}_j^{(\nu)} = \boldsymbol{0} \tag{4.10}$$

for $\nu=1,\ldots,m-1$ and $j=0,\ldots,\tilde{q}-1$.

Next, we satisfy the remaining conditions (2.12) for $\nu=1,\ldots,m-1$, which are

$$\sum_{s=0}^{j}\binom{j}{s}M_{\mathrm{new},j-s}^{(\nu)}\tilde{\boldsymbol{y}}^{(s)} = \boldsymbol{0}. \tag{4.11}$$

Substitute (4.5) to obtain

$$\begin{aligned}-\sum_{s=0}^{j}\binom{j}{s}M_{j-s}^{(\nu)}\tilde{\boldsymbol{y}}^{(s)} &= \sum_{s=0}^{j}\sum_{\ell=0}^{j-s}\binom{j}{s}\binom{j-s}{\ell}m^\ell\Lambda_\ell^{(\nu)}M_{j-s-\ell}^{(0)}\tilde{\boldsymbol{y}}^{(s)}\\ &= \sum_{\ell=0}^{j}\sum_{s=0}^{j-\ell}\binom{j}{\ell}\binom{j-\ell}{s}m^\ell\Lambda_\ell^{(\nu)}M_{j-s-\ell}^{(0)}\tilde{\boldsymbol{y}}^{(s)}\\ &= \sum_{\ell=0}^{j}\binom{j}{\ell}m^\ell\Lambda_\ell^{(\nu)}\sum_{s=0}^{j-\ell}\binom{j-\ell}{s}M_{j-s-\ell}^{(0)}\tilde{\boldsymbol{y}}^{(s)}\\ &= m^j\sum_{\ell=0}^{j}\binom{j}{\ell}\Lambda_\ell^{(\nu)}\tilde{\boldsymbol{y}}^{(j-\ell)},\end{aligned} \tag{4.12}$$

or

$$\sum_{\ell=0}^{j}\binom{j}{\ell}\Lambda_\ell^{(\nu)}\tilde{\boldsymbol{y}}^{(j-\ell)} = -\boldsymbol{z}_j^{(\nu)}. \tag{4.13}$$

For each fixed ν, this can be solved by choosing $\Lambda_j^{(\nu)}$ successively for $j=0,\ldots,\tilde{p}-1$ to satisfy

$$\Lambda_j^{(\nu)}\tilde{\boldsymbol{y}}^{(0)} = -\boldsymbol{z}_j^{(\nu)} - \sum_{\ell=0}^{j-1}\binom{j}{\ell}\Lambda_\ell^{(\nu)}\tilde{\boldsymbol{y}}^{(j-\ell)}. \tag{4.14}$$

The solution is not unique, except in the scalar case.

The matrix A in (4.3) is nonsingular for $N = n = \tilde{p}$, so we can always find a trigonometric polynomial $L^{(\nu)}(\xi)$ of length $\tilde{p}$ or less with an arbitrary starting index k_0 and prescribed moments $\Lambda_j^{(\nu)}$, $j = 0, \dots, \tilde{p} - 1$.

We summarize the results of this section in the following theorem.

THEOREM 4.1. *Assume $\mathcal{H}$, $\tilde{\mathcal{H}}$ are biorthogonal masks with approximation orders q, $\tilde{q}$, respectively, with $\mathcal{H}$ satisfying condition E. Then for any $\tilde{p} \geq 1$ it is possible to find trigonometric polynomials $L^{(\nu)}(\xi)$ of length at most $\tilde{p}$, so that the new masks $\mathcal{H}_{new}$, $\tilde{\mathcal{H}}_{new}$ produced by the lifting process* (3.3) *have approximation orders q, $\tilde{p}$, respectively.*

If $\tilde{\mathcal{H}}_{new}$ satisfies condition E, we can follow the first lifting step with a dual lifting step that produces new masks with approximation orders p, $\tilde{p}$, respectively, for any $p \geq 1$.

It is shown in section 6.3.3 below that if $\tilde{\mathcal{H}}$ satisfies condition E, it is always possible to preserve it during the lifting step.

5. A modified approach. In the procedure in the previous section, it is necessary to impose all $\tilde{p}$ conditions, even if the original $\tilde{\mathcal{H}}$ already has some approximation order $\tilde{q}$. A modified algorithm, suggested in the scalar case in [34], can be adapted to the multiwavelet case.

The motivation is the following. As stated in (3.6) above, the effect of lifting on the multiwavelet functions is described by

$$\boldsymbol{\phi}_{\text{new}}^{(\nu)}(x) = \boldsymbol{\phi}^{(\nu)}(x) + \sum_k L_k^{(\nu)} \boldsymbol{\phi}^{(0)}(x - k). \tag{5.1}$$

Sweldens [34] proposes to replace this by

$$\boldsymbol{\phi}_{\text{new}}^{(\nu)}(x) = \boldsymbol{\phi}^{(\nu)}(x) + \sum_k T_k^{(\nu)} \boldsymbol{\phi}^{(\nu)} \left(\frac{x}{m} - k \right), \tag{5.2}$$

since this preserves existing vanishing moment conditions (2.16).

In our setting, this suggestion amounts to choosing

$$L^{(\nu)}(\xi) = T^{(\nu)}(\xi) H^{(\nu)}(\xi) \tag{5.3}$$

for some shorter trigonometric polynomials $T^{(\nu)}$. It is easy to verify directly that this approach will preserve the existing approximation orders for masks.

THEOREM 5.1. *If $\mathcal{H}$, $\tilde{\mathcal{H}}$ are biorthogonal masks, and $\mathcal{H}_{new}$, $\tilde{\mathcal{H}}_{new}$ are produced by lifting with*

$$L^{(\nu)}(\xi) = T^{(\nu)}(\xi) H^{(\nu)}(\xi), \tag{5.4}$$

then $\tilde{\mathcal{H}}_{new}$ has at least the same approximation order as $\tilde{\mathcal{H}}$.

Proof. Assume that $\tilde{\mathcal{H}}$ has approximation order $\tilde{q}$ or, equivalently (see (4.10))

$$\boldsymbol{z}_j^{(\nu)} = \mathbf{0} \tag{5.5}$$

for $\nu = 1, \dots, m - 1$ and $j = 0, \dots, \tilde{q} - 1$.

Differentiate (5.3) and evaluate at $\xi = 0$ to get

$$\Lambda_\ell^{(\nu)} = \sum_{s=0}^{\ell} \binom{\ell}{s} \Upsilon_s^{(\nu)} M_{\ell - s}^{(\nu)}, \tag{5.6}$$

where $\Upsilon_s^{(\nu)}$ are the moments of $T^{(\nu)}(\xi)$. Thus, for $\nu = 1, \ldots, m-1$ and $j = 0, \ldots, \tilde{q}-1$,

$$\begin{aligned}\sum_{\ell=0}^{j} \binom{j}{\ell} \Lambda_{j-\ell}^{(\nu)} \tilde{\boldsymbol{y}}^{(\ell)} &= \sum_{\ell=0}^{j} \sum_{s=0}^{j-\ell} \binom{j}{\ell} \binom{j-\ell}{s} \Upsilon_s^{(\nu)} M_{j-\ell-s}^{(\nu)} \tilde{\boldsymbol{y}}^{(\ell)} \\ &= \sum_{s=0}^{j} \sum_{\ell=0}^{j-s} \binom{j}{s} \binom{j-s}{\ell} \Upsilon_s^{(\nu)} M_{j-s-\ell}^{(\nu)} \tilde{\boldsymbol{y}}^{(\ell)} \\ &= \sum_{s=0}^{j} \binom{j}{s} m^{j-s} \Upsilon_s^{(\nu)} \boldsymbol{z}_{j-s}^{(\nu)} = \mathbf{0} = -\boldsymbol{z}_j^{(\nu)},\end{aligned} \tag{5.7}$$

so (4.13) is satisfied. ∎

The calculations in (5.7) also provide the equations for determining $\Upsilon_j^{(\nu)}$. As in (4.13), we need

$$\begin{aligned}-\boldsymbol{z}_j^{(\nu)} &= \sum_{\ell=0}^{j} \binom{j}{\ell} \Lambda_{j-\ell}^{(\nu)} \tilde{\boldsymbol{y}}^{(\ell)} \\ &= \sum_{s=0}^{j} \binom{j}{s} m^{j-s} \Upsilon_s^{(\nu)} \boldsymbol{z}_{j-s}^{(\nu)} \\ &= \sum_{s=0}^{j-\tilde{q}} \binom{j}{s} m^{j-s} \Upsilon_s^{(\nu)} \boldsymbol{z}_{j-s}^{(\nu)}\end{aligned} \tag{5.8}$$

for $j = \tilde{q}, \ldots, \tilde{p}-1$. These equations can again be solved successively for $\Upsilon_j^{(\nu)}$ and then $T_{k_0+j}^{(\nu)}$, $j = 0, \ldots, \tilde{p}-\tilde{q}$.

The modified algorithm is faster than the original one. However, it frequently results in longer new masks than the original algorithm. This is illustrated by the examples in section 7.

6. Algorithms. The following algorithms are implementations of the procedures outlined in the previous two sections. They can be used to find suitable lifting factors of any desired length, with free parameters explicitly identified.

Assume that $\mathcal{H}$, $\tilde{\mathcal{H}}$ are biorthogonal masks, with $\mathcal{H}$ satisfying condition E.

6.1. Algorithm 1. Given integers $\tilde{p} \geq 1$, $n \geq 1$, k_0 arbitrary, we want to find matrix trigonometric polynomials of length n with starting index k_0

$$L^{(\nu)}(\xi) = \sum_{k=k_0}^{k_0+n-1} L_k^{(\nu)} e^{-ik\xi} \tag{6.1}$$

so that the new dual mask $\tilde{\mathcal{H}}_{\text{new}}$ produced by the lifting process (3.3) has approximation order $\tilde{p}$.

Step 1. Compute the moments

$$M_j^{(\nu)} = \frac{1}{\sqrt{m}} \sum_k k^j h_k^{(\nu)} \tag{6.2}$$

for $\nu = 0, \ldots, m-1$ and $j = 0, \ldots, \tilde{p}-1$.

Step 2. Compute the vectors $\tilde{\boldsymbol{y}}^{(j)}$, $\boldsymbol{z}_j^{(\nu)}$ for $\nu = 1, \dots, m-1$ and $j = 0, \dots, \tilde{p}-1$ from

$$\begin{aligned} \tilde{\boldsymbol{y}}^{(0)} &= M_0^{(0)} \tilde{\boldsymbol{y}}^{(0)}, \\ \tilde{\boldsymbol{y}}^{(j)} &= \left(m^j I - M_0^{(0)} \right)^{-1} \sum_{s=0}^{j-1} \binom{j}{s} M_{j-s}^{(0)} \tilde{\boldsymbol{y}}^{(s)}, \\ \boldsymbol{z}_j^{(\nu)} &= m^{-j} \sum_{\ell=0}^{j} \binom{j}{\ell} M_\ell^{(\nu)} \tilde{\boldsymbol{y}}^{(j-\ell)}. \end{aligned} \tag{6.3}$$

Step 3. Form the matrices

$$A = \begin{pmatrix} k_0^0 I & k_0^1 I & \cdots & k_0^{\tilde{p}-1} I \\ (k_0+1)^0 I & (k_0+1)^1 I & \cdots & (k_0+1)^{\tilde{p}-1} I \\ \vdots & \vdots & \ddots & \vdots \\ (k_0+n-1)^0 I & (k_0+n-1)^1 I & \cdots & (k_0+n-1)^{\tilde{p}-1} I \end{pmatrix}, \tag{6.4}$$

$$Y = \begin{pmatrix} \binom{0}{0} \tilde{\boldsymbol{y}}^{(0)} & \binom{1}{0} \tilde{\boldsymbol{y}}^{(1)} & \binom{2}{0} \tilde{\boldsymbol{y}}^{(2)} & \cdots & \binom{\tilde{p}-1}{0} \tilde{\boldsymbol{y}}^{(\tilde{p}-1)} \\ & \binom{1}{1} \tilde{\boldsymbol{y}}^{(0)} & \binom{2}{1} \tilde{\boldsymbol{y}}^{(1)} & \cdots & \binom{\tilde{p}-1}{1} \tilde{\boldsymbol{y}}^{(\tilde{p}-2)} \\ & & \binom{2}{2} \tilde{\boldsymbol{y}}^{(0)} & \cdots & \binom{\tilde{p}-1}{2} \tilde{\boldsymbol{y}}^{(\tilde{p}-3)} \\ & & & \ddots & \vdots \\ & & & & \binom{\tilde{p}-1}{\tilde{p}-1} \tilde{\boldsymbol{y}}^{(0)} \end{pmatrix}, \tag{6.5}$$

and

$$Z = \begin{pmatrix} \boldsymbol{z}_0^{(1)} & \cdots & \boldsymbol{z}_{\tilde{p}-1}^{(1)} \\ \vdots & \ddots & \vdots \\ \boldsymbol{z}_0^{(m-1)} & \cdots & \boldsymbol{z}_{\tilde{p}-1}^{(m-1)} \end{pmatrix}. \tag{6.6}$$

The equations (4.13) are equivalent to

$$LAY = -Z, \tag{6.7}$$

where L contains the desired coefficients of $L^{(\nu)}(\xi)$

$$L = \begin{pmatrix} L_{k_0}^{(1)} & \cdots & L_{k_0+\tilde{p}-1}^{(1)} \\ \vdots & \ddots & \vdots \\ L_{k_0}^{(m-1)} & \cdots & L_{k_0+\tilde{p}-1}^{(m-1)} \end{pmatrix}. \tag{6.8}$$

Step 4. Perform a singular value decomposition (SVD)

$$AY = U \Sigma V^*. \tag{6.9}$$

Here U is of size $nr \times nr$, Σ is of size $nr \times \tilde{p}$, and V is of size $\tilde{p} \times \tilde{p}$. Let s be the rank of Σ, then

$$\Sigma = \begin{pmatrix} \Sigma_{11} & 0 \\ 0 & 0 \end{pmatrix} \tag{6.10}$$

with Σ_{11} nonsingular and of size $s \times s$.

Substitute the SVD into (6.7), multiply by V on the right, and partition all matrices corresponding to the partitioning of Σ

$$((LU)_1, (LU)_2) \begin{pmatrix} \Sigma_{11} & 0 \\ 0 & 0 \end{pmatrix} = -((ZV)_1, (ZV)_2), \tag{6.11}$$

or

$$\begin{aligned} (LU)_1 \Sigma_{11} &= -(ZV)_1, \\ 0 &= (ZV)_2. \end{aligned} \tag{6.12}$$

If $(ZV)_2 \neq 0$, the system is unsolvable. Go back to the start and increase n. Otherwise, the solution is

$$\begin{aligned} (LU)_1 &= -(ZV)_1 \Sigma_{11}^{-1}, \\ (LU)_2 &= \text{arbitrary}. \end{aligned} \tag{6.13}$$

The general solution is then

$$L = ((LU)_1, (LU)_2) U^*. \tag{6.14}$$

The free parameters are the elements of $(LU)_2$, of which there are $r(nr - s)(m - 1)$.

Step 5. Assemble the $L^{(\nu)}(\xi)$ and perform the lifting.

Step 6. If required, verify that $\tilde{\mathcal{H}}_{\text{new}}$ satisfies condition E or other properties. If necessary, use optimization on the free parameters to satisfy these conditions.

6.2. Algorithm 2. Given integers $\tilde{p} \geq 1$, $n \geq 1$, k_0 arbitrary, we want to find matrix trigonometric polynomials of length n and starting index k_0

$$T^{(\nu)}(\xi) = \sum_{k=k_0}^{k_0+n-1} T_k^{(\nu)} e^{-ik\xi} \tag{6.15}$$

so that the new dual mask $\tilde{\mathcal{H}}_{\text{new}}$ produced by the lifting process (3.3) performed with $L^{(\nu)}(\xi) = T^{(\nu)}(\xi) H^{(\nu)}(\xi)$ has approximation order $\tilde{p}$.

Steps 1 and 2 are the same as in Algorithm 1.

The existing approximation order $\tilde{q}$ of $\tilde{\mathcal{H}}$ is the largest number $\tilde{q}$ for which $\boldsymbol{z}_j^{(\nu)} = \mathbf{0}$ for all $\nu = 1, \ldots, m$ and $j = 0, \ldots, \tilde{q} - 1$.

Step 3. Form the matrices

$$A = \begin{pmatrix} k_0^0 I & k_0^1 I & \cdots & k_0^{\tilde{p}-\tilde{q}-1} I \\ (k_0+1)^0 I & (k_0+1)^1 I & \cdots & (k_0+1)^{\tilde{p}-\tilde{q}-1} I \\ \vdots & \vdots & \ddots & \vdots \\ (k_0+n-1)^0 I & (k_0+n-1)^1 I & \cdots & (k_0+n-1)^{\tilde{p}-\tilde{q}-1} I \end{pmatrix}, \tag{6.16}$$

$$Y^{(\nu)} = \begin{pmatrix} \binom{\tilde{q}}{0} m^{\tilde{q}} \tilde{\boldsymbol{z}}_{\tilde{q}}^{(\nu)} & \binom{\tilde{q}+1}{0} m^{\tilde{q}+1} \tilde{\boldsymbol{z}}_{\tilde{q}+1}^{(\nu)} & \cdots & \binom{\tilde{p}-1}{0} m^{\tilde{p}-1} \tilde{\boldsymbol{z}}_{\tilde{p}-1}^{(\nu)} \\ & \binom{\tilde{q}+1}{1} m^{\tilde{q}} \tilde{\boldsymbol{z}}_{\tilde{q}}^{(\nu)} & \cdots & \binom{\tilde{p}-1}{1} m^{\tilde{p}-2} \tilde{\boldsymbol{z}}_{\tilde{p}-2}^{(\nu)} \\ & & \ddots & \vdots \\ & & & \binom{\tilde{p}-1}{\tilde{p}-\tilde{q}-1} m^{\tilde{q}} \tilde{\boldsymbol{z}}_{\tilde{q}}^{(\nu)} \end{pmatrix}, \tag{6.17}$$

and

$$Z^{(\nu)} = \left(\boldsymbol{z}_{\tilde{q}}^{(\nu)}, \ldots, \boldsymbol{z}_{\tilde{p}-1}^{(\nu)} \right). \tag{6.18}$$

The equations (5.8) are equivalent to the sequence of equations

$$T^{(\nu)} A Y^{(\nu)} = -Z^{(\nu)} \tag{6.19}$$

for $\nu = 1, \ldots, m-1$, where $T^{(\nu)}$ contains the desired coefficients of $T^{(\nu)}(\xi)$

$$T^{(\nu)} = \left(T_{k_0}^{(\nu)}, \ldots, T_{k_0+n-1}^{(\nu)} \right). \tag{6.20}$$

Step 4. This step is repeated for each $\nu = 1, \ldots, m-1$.

Perform the SVD

$$AY^{(\nu)} = U^{(\nu)} \Sigma^{(\nu)} V^{(\nu)*}. \tag{6.21}$$

and proceed as in Step 4 of Algorithm 1.

The free parameters are the elements of $(T^{(\nu)} U^{(\nu)})_2$. Their total number is

$$r \sum_{\nu} (nr - s^{(\nu)}) = nr^2(m-1) - r \sum_{\nu} s^{(\nu)}. \tag{6.22}$$

Steps 5 and 6 are the same as in Algorithm 1.

6.3. Further comments.

6.3.1. About the implementation. In Algorithm 1, it is possible in Step 3 to solve the equations for all ν simultaneously, since Y is independent of ν. This is not possible in Algorithm 2. As Example 2 in section 7 illustrates, the ranks $s^{(\nu)}$ may vary with ν.

In order to obtain dual approximation order $\tilde{p}$, we need to impose conditions on $\Lambda_j^{(\nu)}$, $j = 1, \ldots, \tilde{p}-1$. We can always find a suitable $L^{(\nu)}$ of length $\tilde{p}$, but we want $L^{(\nu)}$ of length n, with $n < \tilde{p}$ in general. Setting the higher moments to 0 does not result in shorter $L^{(\nu)}$. Instead, we incorporate the matrix A into the algorithms, and solve for $L_j^{(\nu)}$ directly.

6.3.2. Choosing n, k_0. Choosing n as small as possible results in the shortest possible new wavelets, which is usually desirable. Since each approximation order condition amounts to r scalar equations, and each $L_j^{(\nu)}$ contains r^2 coefficients, dual approximation order $\tilde{p}$ should require a smallest possible n of

$$n = \operatorname{ceil}(\tilde{p}/r). \tag{6.23}$$

This cannot be guaranteed (there are counterexamples), but (6.23) gives a good estimate. In particular, constant $L^{(\nu)}$, which does not increase the support lengths of the functions, will in general be able to achieve approximation order r already.

Larger n could be used if extra free parameters are desired.

The starting subscript k_0 of $L^{(\nu)}$ affects both the support length and the centering of the new multiwavelet (see formulas (3.6)). Numerical experiments indicate that the new dual wavelets tend to be smoother if k_0 is chosen so that the scaling function and new dual scaling function are approximately centered around the same point (see Example 1(c) in section 7).

The algorithms could easily be generalized to allow different n, k_0 for each ν.

6.3.3. Stability. We now address the question raised in section 2, regarding the stability properties of the new masks $\mathcal{H}_{\text{new}}$, $\tilde{\mathcal{H}}_{\text{new}}$ (see section 2). The complete answer is not known, but we offer some observations.

1. *Stability of $\boldsymbol{\phi}^{(0)}$ is preserved. Stability of $\tilde{\boldsymbol{\phi}}^{(0)}$ is not preserved in general, but we may be able to preserve it by choosing a suitable L (possibly of higher degree).*

The first part is obvious, since $\boldsymbol{\phi}^{(0)}_{\text{new}} = \boldsymbol{\phi}^{(0)}$.

It is shown in [7] (for scalar wavelets) that any polyphase matrix can be completely factored into lifting steps. Since lifting steps are reversible, this means that any polyphase matrix can be converted into any other by multiple lifting and dual lifting steps. Obviously, stability can get lost in the process.

However, we can always preserve condition E for $\tilde{\boldsymbol{\phi}}^{(0)}$, which is a prerequisite for stability. By (3.5), line 3,

$$\tilde{M}^{(0)}_{\text{new},0} = \tilde{M}^{(0)}_0 - \sum_{\nu=1}^{m-1} \Lambda^{(\nu)*}_0 \tilde{M}^{(\nu)}_0. \tag{6.24}$$

If we choose $\Lambda^{(\nu)}(0) = 0$, then $\tilde{M}^{(0)}_0 = \tilde{M}^{(0)}_{\text{new},0}$, and condition E remains valid. This approach may require increasing n.

In most of the numerical examples we tried, condition E was preserved automatically. In the remaining cases, a simple change in the free parameters was sufficient, with no increase in n needed.

We conjecture that stability of $\tilde{\boldsymbol{\phi}}^{(0)}$ can also be preserved, but the necessary additional conditions on L are not known.

2. *Stability of $\boldsymbol{\phi}$ is preserved.*

If L has finite degree, and $\{\boldsymbol{\phi}^{(\nu)}\}$ form a stable completion of $\boldsymbol{\phi}^{(0)}$, so do $\{\boldsymbol{\phi}^{(\nu)}_{\text{new}}\}$. This follows from Proposition 3.1 in [2].

3. *It is not known whether stability of $\tilde{\boldsymbol{\phi}}$ or stability over all levels for $\boldsymbol{\phi}$ and $\tilde{\boldsymbol{\phi}}$ can be preserved.*

6.3.4. Free parameters. Free parameters that occur during the lifting process can be used for numerical optimization. One defines a function that takes the free parameters as input, calculates the new masks produced by a lifting step with these parameters, and then calculates some objective function which is to be maximized or minimized.

In Example 1(c) in section 7, we used the Sobolev smoothness estimate from [18] as the objective function, to produce the smoothest possible new dual scaling functions for given n and k_0.

7. Examples. We illustrate our algorithms with some numerical examples.

Example 1. This example has dilation factor $m = 2$, multiplicity $r = 2$. We start with cubic Hermite splines as the original scaling function [31]. A basic completion to a biorthogonal pair of masks has the symbols

$$H(z) = \frac{1}{16}\begin{pmatrix} 4+8z+4z^2 & 6-6z^2 \\ -1+z^2 & -1+4z-z^2 \\ 8 & 0 \\ 0 & 8 \end{pmatrix},$$
$$\tilde{H}(z) = \frac{1}{4z}\begin{pmatrix} 4z^2 & 0 \\ 0 & 8z^2 \\ -2+4-2z^2 & -1+z^2 \\ 3-3z^2 & 1+4z+z^2 \end{pmatrix}, \tag{7.1}$$

where $z = \exp(-i\xi)$. The original dual approximation order is 0. $\mathcal{H}$ satisfies condition E, $\tilde{\mathcal{H}}$ does not.

(a) Raise the dual approximation order from 0 to 2.

Algorithm 1 can achieve this with $n = 1$, and no free parameters. For $k_0 = 0$, the result is

$$L(z) = \frac{1}{4}\begin{pmatrix} -2 & 15 \\ 0 & -1 \end{pmatrix}, \tag{7.2}$$

$$\begin{aligned} H_{\text{new}}(z) &= \frac{1}{64}\begin{pmatrix} 16+32z+16z^2 & 24-24z^2 \\ -4+4z^2 & -4+16z-4z^2 \\ 9-16z+7z^2 & -27+60z-3z^2 \\ 1-z^2 & 33-4z+z^2 \end{pmatrix}, \\ \tilde{H}_{\text{new}}(z) &= \frac{1}{16z}\begin{pmatrix} -4+8z+12z^2 & -2+2z^2 \\ 33-60z+27z^2 & 16+4z+18z^2 \\ -8+16z-8z^2 & -4+4z^2 \\ 12-12z^2 & 4+16z+4z^2 \end{pmatrix}. \end{aligned} \tag{7.3}$$

The new masks have length 3. Any other choice of k_0 results in longer masks.

Algorithm 2 gives identical results, since $H^{(1)}(z)$ is a multiple of the identity.

(b) Raise the dual approximation order from 2 to 4, starting with the H_{new} from (a).

Algorithm 1 can achieve this with $n = 2$, and no free parameters. The shortest new masks have length 5, for $k_0 = -1$, and are generated by lifting using

$$L(z) = \frac{1}{48z}\begin{pmatrix} -12+12z & -63-117z \\ 2-2z & 9+21z \end{pmatrix}. \tag{7.4}$$

Algorithm 2 requires only $n = 1$, with no free parameters, but the shortest new masks (also for $k_0 = -1$) have length 7. The lifting factor is

$$L(z) = \frac{1}{13824z}\begin{pmatrix} -729+1152z-423z^2 & -729-3996z+135z^2 \\ 81-160z+79z^2 & -567+636z-39z^2 \end{pmatrix}, \tag{7.5}$$

which is produced from

$$T(z) = \frac{1}{216z}\begin{pmatrix} -72 & -81 \\ 10 & -9 \end{pmatrix}. \tag{7.6}$$

(c) Raise the dual approximation order from 0 to 2 (starting again with the original $\mathcal{H}$, $\tilde{\mathcal{H}}$) with free parameters, and optimize for smoothness.

The algorithm described in [18] can be used to determine a lower bound on the Sobolev exponent s of a multiscaling function. The shortest dual scaling function with approximation order 2 derived in (a) is in the Sobolev space $W^{-1.2294}$, so it is not even an L^2-function.

If we apply Algorithm 1 with $n = 2$, there are 4 free parameters. The shortest new scaling function symbols have length 5 for $k_0 = -1$ or $k_0 = 0$.

For $k_0 = 0$, the coefficients of H_{new} and $\tilde{H}_{\text{new}}$ are centered at 2 and -1, respectively. Numerical optimization of the Sobolev exponent yields a smoothest $\tilde{H}_{\text{new}}$ in $W^{-0.7877}$.

For $k_0 = -1$, the centers are 0 and 1, which is a better fit. The smoothest $\tilde{H}_{\text{new}}$ is in $W^{0.8289}$, which matches the result of [30], [32]. This $\tilde{H}_{\text{new}}$ satisfies condition E, and could be used as the basis for a further dual lifting step.

Example 2. This example has dilation factor $m = 3$, multiplicity $r = 1$. We take $\phi^{(0)}$ to be the characteristic function of $[0, 1]$, i.e., the Haar scaling function, with approximation order 1. A completion with dual approximation order 1 is

$$H(z) = \frac{1}{9}\begin{pmatrix} 3 + 3z + 3z^2 \\ \sqrt{3}\left(-1 + 2z - z^2\right) \\ \sqrt{3}\left(-1 - z + 2z^2\right) \end{pmatrix},$$
$$\tilde{H}(z) = \frac{1}{3}\begin{pmatrix} 1 + z + z^2 \\ \sqrt{3}\left(-1 + z\right) \\ \sqrt{3}\left(-1 + z^2\right) \end{pmatrix}. \tag{7.7}$$

Condition E is satisfied by the original masks. Since $r = 1$, it is automatically preserved.

We want to raise the dual approximation order to 3.

Algorithm 1 requires $n = 3$, with no free parameters. The choice $k_0 = -1$ produces the shortest new masks of length 9

$$H_{\text{new}}(z) = \frac{1}{243z^3}\begin{pmatrix} 81z^3 + 81z^4 + 81z^5 \\ \sqrt{3}\left(1 + z + z^2 - 29z^3 + 52z^4 - 29z^5 + z^6 + z^7 + z^8\right) \\ \sqrt{3}\left(4 + 4z + 4z^2 - 26z^3 - 26z^4 + 55z^5 - 5z^6 - 5z^7 - 5z^8\right) \end{pmatrix},$$
$$\tilde{H}_{\text{new}}(z) = \frac{1}{81z^3}\begin{pmatrix} -4 - z + 5z^2 + 26z^3 + 29z^4 + 26z^5 + 5z^6 - z^7 - 4z^8 \\ \sqrt{3}\left(-27z^3 + 27z^4\right) \\ \sqrt{3}\left(-27z^3 + 27z^5\right) \end{pmatrix}, \tag{7.8}$$

using

$$L(z) = \frac{1}{27\sqrt{3}z}\begin{pmatrix} 1 - 2z + z^2 \\ 4 + z - 5z^2 \end{pmatrix}. \tag{7.9}$$

Algorithm 2 requires at least $n = 2$, and produces new masks of length 12. There is one free parameter, in $T^{(1)}$ only.

Acknowledgments. This article was written during the author's sabbatical leave at Flinders University, Adelaide, South Australia. It is part of a long-term research project involving the design of wavelets with prescribed properties, directed by Prof. Jaroslav Kautsky. I would like to thank Profs. Kautsky and Bill Moran for their hospitality and support during this time, and to thank them and other group members (notably Radka Turcajová and Vasily Strela) for inspiration and helpful discussions.

I would also like to thank the anonymous referees for helpful suggestions, in particular for a simplified proof of Theorem 2.1 and clarifications on stability questions.

REFERENCES

[1] B. K. Alpert, *Sparse Representation of Smooth Linear Operators*, dissertation, Yale University, New Haven, CT, 1990.

[2] J. M. Carnicer, W. Dahmen, and J. M. Peña, *Local decomposition of refinable spaces and wavelets*, Appl. Comput. Harmon. Anal., 3 (1996), pp. 127–153.

[3] D.-R. Chen, B. Han, and S. D. Riemenschneider, *Construction of multivariate biorthogonal wavelets with arbitrary vanishing moments*, Adv. Comput. Math., 13 (2000), pp. 131–165.
[4] A. Cohen, I. Daubechies, and G. Plonka, *Regularity of refinable function vectors*, J. Fourier Anal. Appl., 3 (1997), pp. 295–324.
[5] W. Dahmen, *Decomposition of refinable spaces and applications to operator equations*, Numer. Algorithms, 5 (1993), pp. 229–245.
[6] W. Dahmen and C. A. Micchelli, *Biorthogonal wavelet expansions*, Constr. Approx., 13 (1997), pp. 293–328.
[7] I. Daubechies and W. Sweldens, *Factoring wavelet transforms into lifting steps*, J. Fourier Anal. Appl., 4 (1998), pp. 247–269.
[8] G. M. Davis, V. Strela, and R. Turcajová, *Multiwavelet construction via the lifting scheme*, in Wavelet Analysis and Multiresolution Methods, T.-X. He, ed., Lecture Notes in Pure and Appl. Math., Marcel Dekker, New York, 1999.
[9] J. S. Geronimo, D. P. Hardin, and P. R. Massopust, *Fractal functions and wavelet expansions based on several scaling functions*, J. Approx. Theory, 78 (1994), pp. 373–401.
[10] T. N. T. Goodman, *Construction of wavelets with multiplicity*, Rend. Mat. Appl. (7), 14 (1994), pp. 665–691.
[11] T. N. T. Goodman and S. L. Lee, *Wavelets of multiplicity r*, Trans. Amer. Math. Soc., 342 (1994), pp. 307–324.
[12] T. N. T. Goodman, S. L. Lee, and W. S. Tang, *Wavelets in wandering subspaces*, Trans. Amer. Math. Soc., 338 (1993), pp. 639–654.
[13] B. Han, *Analysis and construction of optimal multivariate biorthogonal wavelets with compact support*, SIAM J. Math. Anal., 31 (1999), pp. 274–304.
[14] C. Heil and D. Colella, *Matrix refinement equations: Existence and uniqueness*, J. Fourier Anal. Appl., 2 (1996), pp. 363–377.
[15] C. Heil, G. Strang, and V. Strela, *Approximation by translates of refinable functions*, Numer. Math., 73 (1996), pp. 75–94.
[16] P. N. Heller, *Rank M wavelets with N vanishing moments*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 502–519.
[17] R.-Q. Jia, S. D. Riemenschneider, and D. X. Zhou, *Approximation by multiple refinable functions*, Canad. J. Math., 49 (1997), pp. 944–962.
[18] Q. Jiang, *On the regularity of matrix refinable functions*, SIAM J. Math. Anal., 29 (1998), pp. 1157–1176.
[19] Q. Jiang and Z. Shen, *On existence and weak stability of matrix refinable functions*, Constr. Approx., 15 (1999), pp. 337–353.
[20] J. Kovačević and W. Sweldens, *Wavelet families of increasing order in arbitrary dimensions*, IEEE Trans. Image Process., 9 (2000), pp. 480–496.
[21] S. G. Mallat, *Multiresolution approximations and wavelet orthonormal bases of $L^2(R)$*, Trans. Amer. Math. Soc., 315 (1989), pp. 69–87.
[22] G. Plonka, *Approximation properties of multiscaling functions: A Fourier approach*, Rostock. Math. Kolloq., 49 (1995), pp. 115–126.
[23] G. Plonka, *Approximation order provided by refinable function vectors*, Constr. Approx., 13 (1997), pp. 221–244.
[24] G. Plonka and V. Strela, *Construction of multiscaling functions with approximation and symmetry*, SIAM J. Math. Anal., 29 (1998), pp. 481–510.
[25] G. Plonka and V. Strela, *From wavelets to multiwavelets*, in Mathematical Methods for Curves and Surfaces II, M. Daehlen, T. Lyche, and L. L. Schumaker, eds., Vanderbilt University Press, Nashville, TN, 1998, pp. 1–25.
[26] Z. Shen, *Refinable function vectors*, SIAM J. Math. Anal., 29 (1998), pp. 235–250.
[27] P. Steffen, P. N. Heller, R. A. Gopinath, and C. S. Burrus, *Theory of regular M-band wavelet bases*, IEEE Trans. Signal Process., 41 (1993), pp. 3497–3511.
[28] V. Strela, *Multiwavelets: Theory and Applications*, Ph.D. thesis, MIT, Cambridge, MA, 1996.
[29] V. Strela, *Multiwavelets: Regularity, orthogonality and symmetry via two-scale similarity transform*, Studies Appl. Math., 98 (1997), pp. 335–354.
[30] V. Strela, *A note on construction of biorthogonal multi-scaling functions*, in Wavelets, Multiwavelets, and Their Applications, Contemp. Math., A. Aldroubi and E. B. Lin, eds., Amer. Math. Soc., Providence, RI, 1998, pp. 149–157.
[31] V. Strela and G. Strang, *Finite element multiwavelets*, in Proceedings of the Maratea NATO Conference, NATO Adv. Sci. Inst. Ser. C. Math. Phys. Sci. 454, Kluwer, Dordrecht, The Netherlands, 1995.
[32] V. Strela and R. Turcajová, *Two ways of constructing Hermite spline multiwavelets*, in preparation.

[33] W. Sweldens, *The lifting scheme: A custom-design construction of biorthogonal wavelets*, Appl. Comput. Harmon. Anal., 3 (1996), pp. 186–200.
[34] W. Sweldens, *The lifting scheme: A construction of second generation wavelets*, SIAM J. Math. Anal., 29 (1998), pp. 511–546.
[35] P. P. Vaidyanathan, *Quadrature mirror filter banks, M-band extensions and perfect-reconstruction techniques*, IEEE ASSP Magazine, 4 (1987), pp. 4–20.
[36] M. Vetterli and C. Herley, *Wavelets and filter banks: Theory and design*, IEEE Trans. Acoust. Speech Signal Process., 40 (1992), pp. 2207–2232.

SIAM J. MATH. ANAL.
Vol. 32, No. 5, pp. 1050–1080

LARGE TIME ASYMPTOTICS OF SOLUTIONS AROUND SOLITARY WAVES TO THE GENERALIZED KORTEWEG–DE VRIES EQUATIONS*

TETSU MIZUMACHI†

Abstract. We consider the long time asymptotics of solutions that are close to a solitary wave solution to the generalized Korteweg–de Vries equation

$$u_t + u^p u_x + u_{xxx} = 0 \quad \text{for } x \in \mathbb{R},\ t > 0.$$

If $1 \le p < 4$ and the spectrum of the linearized equation around the initial solitary wave has the simplest possible structure, the solitary wave turns out to be asymptotically stable with respect to finite energy perturbations with polynomial decay as $x \to \infty$. Furthermore, we show that the asymptotics of the solution for large time is given by a sum of a solitary wave with slightly displaced parameters and a small dispersion if $2 < p < 4$.

Key words. generalized Korteweg–de Vries equation, solitary waves, nonlinear scattering

AMS subject classifications. 35Q53, 35B40, 76B25

PII. S0036141098346827

1. Introduction.

In the present paper, we study the large time behavior of solutions around solitary waves to the generalized Korteweg–de Vries (GKdV) equation

$$u_t + f(u)_x + u_{xxx} = 0 \quad \text{for } x \in \mathbb{R},\ t > 0, \tag{1.1}$$

$$u(x,0) = u_0(x) \quad \text{for } x \in \mathbb{R}, \tag{1.2}$$

where

$$f(u) = u^{p+1}/(p+1).$$

For $p = 1$, the equation was derived by Korteweg and de Vries in [23] as a model for long waves propagating in a canal.

Solitary wave solutions are a class of spatially localized solutions with finite energy. The GKdV equation has a two-parameter family of solitary wave solutions of the form $u(x,t) = \varphi_c(x - ct + \gamma)$, where c is a positive number, $\gamma \in \mathbb{R}$, and $\varphi_c(y)$ is a positive symmetric function to

$$\varphi_c'' - c\varphi_c + f(\varphi_c) = 0 \quad \text{for } y \in \mathbb{R}.$$

Now, let

$$u_0(x) = \varphi_{c_0}(x + \gamma_0) + v_0(x), \tag{1.3}$$

where $c_0 > 0$, $\gamma_0 \in \mathbb{R}$, and $v_0 \in H^1(\mathbb{R})$. The orbital stability of the solitary wave solutions has been studied by Benjamin [3], Bona [4], Bona, Souganidis, and Strauss [8], and Weinstein [37] (see also Bona and Soyeur [9]). They proved that the solitary

*Received by the editors October 29, 1998; accepted for publication (in revised form) September 13, 2000; published electronically January 31, 2001.
http://www.siam.org/journals/sima/32-5/34682.html

†Department of Mathematical Sciences, Yokohama City University, 22-2 Seto, 236-0027, Yokohama, Japan (Tetsu.Mizumachi@math.yokohama-cu.ac.jp). This research is supported by Research Fellowships of the Japan Society for the promotion of Science for Young Scientists.

wave solutions are stable if $0 < p < 4$ and unstable if $p \geq 4$. That is to say, the solution u to (1.1)–(1.3) with small v_0 remains close to the set $\{\varphi_{c_0}(x - c_0 t + \gamma) \,|\, \gamma \in \mathbb{R}\}$ for all the time if $0 < p < 4$.

If $p = 1$ or 2, the inverse scattering theory is available. It informs us that the solution to (1.1) with well-localized initial data resolves into a train of solitary waves moving to the right and dispersive radiation which moves to the left (see [1], [12], [13], [14], [31]). Although the inverse scattering theory does not apply to (1.1) with more general p, there is some numerical evidence that shows that this type of asymptotic resolution extends to equations with more general nonlinearities (see [5], [6], [7]).

Pego and Weinstein [28] proved the asymptotic stability of solitary waves with an exponential spatial weight e^{ay} $(a > 0)$ in the case where $p = 1$, 2 or $3 \leq p < 4$ and the linearized operator around solitary waves has no eigenvalue in L^2 other than 0.

However, exponential localization seems to be a strong constraint. Indeed, if there exists a small soliton other than the main wave, their results cannot be applied because solitons are not small in $H^1 \cap H^1_a$, where $H^1_a = \{v \,|\, e^{ay} v \in H^1\}$ with $\|v\|_{H^1_a} = \|e^{ay} v\|_{H^1}$ (see [28, p. 337]).

To deal with a more general class of perturbations, it is natural to use the algebraically weighted space. Recently, Miller [24] has shown some estimates of the solutions to the linearized equation in algebraically weighted spaces following the lines of Jensen and Kato [18].

One of the purposes of the present paper is to show the asymptotic stability of solitary waves in algebraically weighted spaces. Making use of the weighted L^p-L^q estimate, which is a generalization of the linear estimates in [24], we prove the asymptotic stability of solitary waves in H^1 with a weight function growing polynomially as $x \to \infty$ and decaying exponentially as $x \to -\infty$ (see Theorem 2.2 in section 2). Our result can deal with the case where the small soliton is behind and well apart from the dominant soliton. Furthermore, our result covers the case where $1 < p < 3$, which was left open in [28].

The other purpose of the present paper is to show the existence of scattering states around solitary wave solutions of the GKdV equation. There is extensive literature on the nonlinear scattering of solutions to (1.1)–(1.2) if $u_0(x)$ is small (see [11], [16], [20], [21], [22], [29], [30], [32], [35], and the references therein). They tell us that small solutions to (1.1) decay at the same rate as solutions to the linear problem

$$u_t + u_{xxx} = 0 \quad \text{for } x \in \mathbb{R} \text{ and } t > 0 \tag{1.4}$$

and are asymptotically free if $p > 2$. There is also a conjecture in [30] that the small solutions are not asymptotically free for $1 < p \leq 2$.

On the other hand, the existence of scattering states around standing wave solutions has been studied for some classes of nonlinear Schrödinger equation (see [10], [33], [34]). To the best of our knowledge, however, the corresponding result for the GKdV equation remains unknown. We prove that the dispersive wave part of the solution around the solitary wave also behaves like the free dispersion wave as $t \to \infty$ if $p > 2$ (see Theorem 2.4 in section 2).

To prove the result, we show that the interaction of the dispersive part and the solitary wave becomes small in various norms as $t \to \infty$, and we apply the method due to Hayashi and Naumkin [16], which shows nonlinear scattering of solutions to the GKdV equation in the case where $p > 2$ and the initial data are small.

The plan of the present paper is as follows. In section 2, we introduce our main results and several lemmas which shall be used in the sections that follow. In section 3,

we will show that the solution u to (1.1)–(1.3) which is close to solitary waves in an algebraically weighted space can be expressed in the form

$$u(x,t) = \varphi_{c(t)}(y,t) + v(y,t) \quad \text{with } y = x - \int_0^t c(s)ds + \gamma(t),$$

following the lines of [28], [10], [33], and [34]. Here the parameters of speed $c(t)$ and phase shift $\gamma(t)$ vary in time so that $\varphi_{c(t)}(y)$ describes the motion of the solitary wave part for each t. In section 4, we will derive some a priori estimates of (v, c, γ) and show that the solution which is initially close to a solitary wave tends to a nearby solitary wave as $t \to \infty$ in a local sense relative to a frame moving with the solitary wave. In section 5, we show that solutions which are close to a solitary wave in some weighted space can be resolved into a solitary wave and a purely dispersive wave. In the appendix, we estimate the rate of decay of solutions to the linearized equation in some weighted spaces (see Lemma 2.7 and Corollaries 2.9 and 2.10).

Finally, let us introduce several notations, which shall be used later. Let $x_+ = \max(x, 0)$, $x_- = \max(-x, 0)$, $\langle x \rangle = (1 + x^2)^{1/2}$ for $x \in \mathbb{R}$. We use the notations $\|\cdot\|_p$ for the $L^p(\mathbb{R})$-norm and $\|\cdot\|_{m,s}$ for the norms defined by $\|v\|_{m,s} = \|\langle x \rangle^s (1-\partial^2)^{m/2} v\|_2$. For simplicity, we denote by $\|\cdot\|$ and $(\cdot,\cdot)$ the norm and the inner product of $L^2(\mathbb{R})$, respectively. For $\alpha > 0$ and $b > 0$, let $h_{\alpha,b}(x)$ be a smooth function that satisfies $h_{\alpha,b}(x) > 0$, $h'_{\alpha,b} > 0$ for $x \in \mathbb{R}$, $\sup_{x\in\mathbb{R}} |h^{(k)}_{\alpha,b}(x)/h_{\alpha,b}(x)| < \infty$ for $k \in \mathbb{N} \cup \{0\}$, and

$$h^{(k)}_{\alpha,b}(x) = \begin{cases} O(|x|^{\alpha-k}) & \text{as } x \to \infty, \\ O(e^{bx}) & \text{as } x \to -\infty \end{cases}$$

for $0 \le k \le [\alpha]$, where $h^{(k)}_{\alpha,b}(x)$ denotes the kth derivative of $h_{\alpha,b}(x)$. For the definiteness, we choose a nonnegative C^∞-function $\varphi(x) \in C_0^\infty[-1, 1]$ with $\|\varphi\|_1 = 1$ and define $h_{\alpha,b}$ by $h_{\alpha,b}(x) = \exp\left\{\int_0^x \omega_{\alpha,b} * \varphi(y)dy\right\}$ for $\alpha > 0$, $b > 0$, where

$$\omega_{\alpha,b}(x) = \begin{cases} \alpha & \text{for } x \le 1, \\ b(x - 1 + b/\alpha)^{-1} & \text{for } x \ge 1. \end{cases}$$

This family of functions meets all our requirements and will be used without further comments. For $\alpha < 0$ and $b < 0$, we define $h_{\alpha,b}(x)$ by $h_{\alpha,b}(x) = h_{-\alpha,-b}(x)^{-1}$. We denote by $L^p(\alpha, b)$, $W^{k,p}(\alpha, b)$, and $L^{s,p}(\alpha, b)$ the weighted spaces with the norms defined by

$$\|v\|_{L^p(\alpha,b)} = \|h_{\alpha,b} v\|_p, \quad \|v\|_{W^{k,p}(\alpha,b)} = \left(\sum_{|\beta|\le k} \|h_{\alpha,b} \partial_x^\beta v\|_p^p \right)^{1/p},$$

$$\|v\|_{L^{s,p}(\alpha,b)} = \|(1 - \partial_x^2)^{s/2} h_{\alpha,b} v\|_p,$$

where $k \in \mathbb{N} \cup \{0\}$, $p \in [1, \infty]$, $s \in \mathbb{R}$. Note that $W^{k,p}(\alpha, b)$ and $L^{k,p}(\alpha, b)$ are equivalent if $1 < p < \infty$ and that the space $L^{-s,p'}(-\alpha, -b)$ is the dual space of $L^{s,p}(\alpha, b)$ for $1 < p < \infty$, where $1/p + 1/p' = 1$.

We define D^α as

$$D^\alpha f \equiv \mathcal{F}^{-1} \xi^\alpha e^{-i\pi(1+\alpha)/2} \mathcal{F} f$$

$$= \frac{2\pi}{\Gamma(1-\alpha)} \int_0^\infty (f(x+y) - f(x)) \frac{dy}{y^{\alpha+1}}$$

for $\alpha \in (0,1)$, where $\mathcal{F}f(\xi) = \int e^{-ix\xi} f(x)dx$ and $\mathcal{F}^{-1}g(x) = \frac{1}{2\pi}\int e^{ix\xi} g(\xi)d\xi$. In the course of calculations, various constants will be denoted simply by C and C_i, which are possibly different from one line to the next.

2. Assumptions and results. First, we recall the spectral properties of the linearized operator of (1.1) around the solitary wave solutions. Let $A_{[c]}$ be the operator defined by

$$A_{[c]}u \equiv -\partial_y\{\partial_y^2 - c + f'(\varphi_c(y))\}u. \tag{2.1}$$

Obviously, the essential spectrum of the linearized operator $A_{[c]}$ consists of $i\mathbb{R}$ and it holds that $A_{[c]}\partial_y\varphi_c = 0$, $A_{[c]}\partial_c\varphi_c = -\partial_y\varphi_c$. So $\lambda = 0$ is always an eigenvalue of $A_{[c]}$ embedded in the essential spectrum. Put

$$\xi_1(y,\, c) = \partial_y\varphi_c(y), \quad \xi_2(y,\, c) = \partial_c\varphi_c(y),$$
$$\eta_1(y,c) = \theta_1\int_{-\infty}^{y}\partial_c\varphi_c + \theta_2\varphi_c(y), \quad \eta_2(y,c) = \theta_3\varphi_c(y),$$

where

$$\theta_3 = -\theta_1 = 2\left(\frac{d}{dc}\|\varphi_c\|^2\right)^{-1} \quad \text{and} \quad \theta_2 = 2\left(\frac{d}{dc}\int_{\mathbb{R}}\varphi_c\right)^2\left(\frac{d}{dc}\|\varphi_c\|^2\right)^{-2}. \tag{2.2}$$

The functions $\xi_1(y,c)$, $\xi_2(y,c)$, and $\eta_2(y,c)$ decay exponentially as $|y| \to \infty$. The function $\eta_1(y)$ also decays exponentially as $y \to -\infty$, but it is merely bounded as $y \to \infty$. In addition, it holds that

$$A_{[c]}\xi_1(y,c) = 0, \quad A_{[c]}\xi_2(y,c) = -\xi_1(y,c), \tag{2.3}$$
$$A_{[c]}^*\eta_1(y,c) = -\eta_2(y,c), \quad A_{[c]}^*\eta_2(y,c) = 0, \tag{2.4}$$
$$\langle\xi_i(\cdot,\, c),\, \eta_j(\cdot,\, c)\rangle = \delta_{ij} \quad \text{for } i,j = 1,2, \tag{2.5}$$

where $\langle\cdot,\cdot\rangle$ is defined by $\langle f,g\rangle = \int f\bar{g}dx$.

Let P_c and Q_c be the projections defined by

$$P_cu = \sum_{i=1}^{2}\langle u,\, \eta_i(\cdot,c)\rangle\xi_i(\cdot,c) \quad \text{and} \quad Q_cu = (1-P_c)u \tag{2.6}$$

in $W^{k,p}(\alpha,b)$. The operators P_c and Q_c are well defined if $1 \le p \le \infty$, $\alpha > 1 - 1/p$, and $b > 0$ is a sufficiently small number.

For the operator $A_{[c]}$ defined in L^2, put

$$\mathrm{Ker}(A_{[c]}) = \{w \in D(A_{[c]}) \mid A_{[c]}u = 0\}, \quad \mathrm{Ker}_g(A_{[c]}) = \bigcup_{k=0}^{\infty}\mathrm{Ker}(A_{[c]}^k).$$

The following proposition due to Pego and Weinstein [27], [28] tells us that $A_{[c]}$ generically has no eigenvalue in $L^2(\mathbb{R})$ other than 0.

PROPOSITION 2.1 (see Pego and Weinstein [27], [28]).

(i) *Assume $0 < p \le 4$. Then $A_{[c]}$ has no isolated eigenvalues. Its spectrum coincides with the imaginary axis.*

(ii) *Assume* $\frac{d}{dc}\|\varphi_c\|_{L^2}^2 \neq 0$ $(p \neq 4)$. *Then* $\lambda = 0$ *is an eigenvalue for* $A_{[c]}$ *with algebraic multiplicity two. More precisely, it holds that*

$$\mathrm{Ker}_g(A_{[c]}) = \mathrm{Ker}(A_{[c]}^2) = \mathrm{span}\{\xi_1(\cdot, c), \xi_2(\cdot, c)\},$$
$$\mathrm{Ker}_g(A_{[c]}^*) = \mathrm{Ker}(A_{[c]}^{*2}) = \mathrm{span}\{\eta_1(\cdot, c), \eta_2(\cdot, c)\},$$

where $\mathrm{span}\{w_1, w_2\}$ *denotes the linear subspace* $\{\alpha w_1 + \beta w_2 \,|\, \alpha, \beta \in \mathbb{C}\}$.

(iii) *The set* $\mathbf{E}$*, of values* p *with* $p > 0$*, where* $A_{[c]}$ *has a nonzero eigenvalue on the imaginary axis, is a discrete set. In particular,* $\mathbf{E} \cap [1, 4]$ *is a finite set, which does not include the values* $p = 1$ *and* $p = 2$.

We will assume the absence of nonzero eigenvalues of $A_{[c]}$ in the closed right half plane of $\mathbb{C}$ throughout the paper.

Now, we are in position to state our main results.

THEOREM 2.2 (asymptotic stability). *Assume* $1 \leq p < 4$ *and* $p \notin \mathbf{E}$*, where* $\mathbf{E}$ *is a set defined in Proposition* 2.1. *Let* u *be the solution to* (1.1)–(1.3). *Let* α_1, α_2 *be positive numbers such that* $\alpha_1 = \alpha_2 + r + 1$, $\alpha_2 = r + 1/2$, *and* $r \geq 2$. *Let* b *be a sufficiently small positive number. Then, there exist some* $\varepsilon_0 > 0$ *and* $C > 0$ *satisfying the following. If*

$$\|v_0\|_{1,0} + \|v_0\|_{L^2(\alpha_1, b)} = \varepsilon$$

for $0 < \varepsilon < \varepsilon_0$*, there exists* $(c_+, \gamma_+) \in \mathbb{R}_+ \times \mathbb{R}$ *such that*

$$|c_+ - c_0| + |\gamma_+ - \gamma_0| < C\varepsilon, \tag{2.7}$$
$$\sup_{t \geq 0} \|u(\cdot, t) - \varphi_{c_+}(\cdot - c_+ t + \gamma_+)\|_{1,0} \leq C\varepsilon, \tag{2.8}$$
$$\sup_{t \geq 0} t^r \|u(\cdot + c_+ t - \gamma_+, t) - \varphi_{c_+}(\cdot)\|_{W^{1,2}(\alpha_2, b)} \leq C\varepsilon. \tag{2.9}$$

Remark 2.3. It is expected that more or less initial data u_0 evolve, in a fairly short time, into a train of solitary waves plus a dispersive trail. If $p = 1$ or 2, the phenomenon is understood analytically as a consequence of inverse scattering transform. Although the inverse scattering theory is not available in the nonintegrable case, there is some numerical evidence which suggests that the phenomenon extends for more general p (see, for example, [5], [6], [7]).

Pego and Weinstein [28] proved the asymptotic stability of the main wave in an exponentially weighted space. That is, if $u_0(x) = \varphi_{c_0}(x + \gamma_0) + v_0(x)$ and $\|v_0\|_{H^1} + \|e^{ax} v_0\|_{H^1}$ is sufficiently small for an $a > 0$, the solution arising from such initial data tends toward a nearby solitary wave in the weighted space. Although their method is broadly useful to investigate the large time asymptotics of nonintegrable equations, their result does not apply to multipulse solutions for the following reason. Let c_1, c_2, γ_1, and γ_2 be numbers satisfying $c_1 > c_2 > 0$ and $\gamma_1 < \gamma_2$. Suppose that u is a 2-soliton solution to the KdV equation that approaches $\varphi_{c_1}(x - c_1 t + \gamma_1) + \varphi_{c_2}(x - c_2 t + \gamma_2)$ as $t \to \infty$. Then the solution $u(x, t)$ spatially decays at the same rate as $e^{-\sqrt{c_2} x}$ as $x \to \infty$. So the assumption $e^{ax} v_0 \in H^1$ imposes the minimum size on the amplitude of solitary wave in the combination, since we must have $\sqrt{c_2} > a$. On the other hand, since $\varphi_c(x) = c^{1/p} \varphi_1(c^{1/2} x)$, we have $c \to 0$ as $\|\varphi_c\|_{H^1} \to 0$. But there is no guarantee that the H^1 norm of φ_{c_2} with $\sqrt{c_2} > a$ is enough to regard u as a small perturbation to $\varphi_{c_1}(x - c_1 t + \gamma_1)$ and apply their argument.

To deal with more general classes of perturbations from which solitary waves arise, we make use of the weight function that grows polynomially as $x \to \infty$. Since

$\|\varphi_c(\cdot+\gamma)\|_{L^2(\alpha_1,b)} \to 0$ as $c \to 0$ and $\gamma \to 0$, we can regard an N-soliton solution as a perturbation to the dominant solitary wave if the $N-1$ of the waves are small and far behind the main wave.

The following theorem shows that the initial data which is close to a solitary wave (in a certain weighted space) resolves into a solitary wave with shifted parameters plus the linear evolution.

THEOREM 2.4 (scattering). *Let $2 < p < 4$, $p \notin \mathbf{E}$, and $\alpha > 11/2$. Suppose that*

$$\|(1+x_+)^\alpha v_0\| + \|v_0\|_{1,1} = \varepsilon$$

is sufficiently small. Then, there exists $V \in L^2(\mathbb{R})$ and (c_+, γ_+) satisfying (2.7)–(2.9) *with $\alpha_2 = \alpha/2 - 1/4$ and*

$$\|u(\cdot,t) - \varphi_{c_+}(\cdot - c_+t + \gamma_+) - e^{-t\partial_x^3}V\| = o(1) \tag{2.10}$$

as $t \to \infty$.

Remark 2.5. Noting that $\varphi_c(x) = c^{1/p}\varphi_1(c^{1/2}x)$, we see that $\|\varphi_c\|_{1,1}$ does not tend to 0 as $c \to 0$. So the smallness of v_0 in $H^{1,1}$ precludes the possibility of the emergence of small solitons ahead or behind the main wave.

Remark 2.6. The solitary wave solutions are unstable if $p \geq 4$ (see [8]). There are some numerical experiments that show that solutions near the solitary wave blow up in a finite time (see [6], [7]).

On the other hand, if $p \leq 2$, the dispersive wave part is expected not to be asymptotically free as $t \to \infty$ (see [30]).

To prove the above theorems, we need some results on the local decay of solutions to the linearized equation. Let $u(t,x)$ be the solution to

$$\begin{cases} \partial_t u - A_{[c]}u = 0, \\ u(0,x) = u_0(x), \end{cases}$$

and let $U_c(t)u_0 := u(t,x)$. Then we have the following lemma.

LEMMA 2.7. *Assume that $1 \leq p < 4$ and that $p \notin \mathbf{E}$. Let $1 \leq q_1 \leq q_2 < \infty$ and let b be a sufficiently small number. Let $k \in \mathbb{N} \cup \{0\}$, $m = 0,1,2,3$, and let r be a real number with $r \geq [(m+1)/2] + 1$. Suppose that $\alpha_1 \geq \alpha_2 + r + 1 + 1/q_2 - 1/q_1$, $\alpha_2 > 1 - 1/q_2$. Then there exists a $C_1 > 0$ such that*

$$\|U_c(t)Q_cu_0\|_{W^{k+m,q_2}(\alpha_2,b)} \leq C_1 t^{-r}\|Q_cu_0\|_{W^{k,q_1}(\alpha_1,b)}$$

for every $u_0 \in W^{k,q_1}(\alpha_1,b)$ and $t > 0$. Especially, there exists a $C_2 > 0$ such that

$$\|U_c(t)Q_cu_0\|_{W^{k,2}(\alpha_2,b)} \leq C_2(1+t)^{-r}\|Q_cu_0\|_{W^{k,2}(\alpha_1,b)}$$

for every $u_0 \in W^{k,2}(\alpha_1,b)$ and $t \geq 0$. Furthermore, if I is a compact subset of $(0,\infty)$, the constants C_1 and C_2 can be chosen uniformly for all $c \in I$.

Remark 2.8. Lemma 2.7 of this paper is a generalization of Theorem 1.1 in [24]. The difference between our Lemma 2.7 and Theorem 1.1 of [24] is that q_1 need not be equal to q_2, which enables us to prove the asymptotic stability in the algebraically weighted space.

COROLLARY 2.9. *Let $1 \leq q_1 \leq q_2 \leq \infty$ and $q_1 \neq \infty$. Let p, α_1, α_2, and b be as in Lemma* 2.7. *Let θ be a number satisfying*

$$\theta = \begin{cases} 1/4 & \text{if } 1/q_1 - 1/q_2 < 3/4, \\ 1/3 & \text{if } 1/q_1 - 1/q_2 \geq 3/4. \end{cases}$$

If $r \geq 1$, there exists a $C_1 > 0$ such that

$$\|U_c(t)Q_c u_0\|_{L^{q_2}(\alpha_2,b)} \leq C_1 t^{-\theta}(1+t)^{-r+\theta}\|Q_c u_0\|_{L^{q_1}(\alpha_1,b)}$$

for every $t > 0$ and $u_0 \in L^{q_1}(\alpha_1, b)$. If $r \geq 2$, there exists a $C_2 > 0$ such that

$$\|\partial_x U_c(t)Q_c u_0\|_{L^{q_2}(\alpha_2,b)} \leq C_2 t^{-3/4}(1+t)^{-r+3/4}\|Q_c u_0\|_{L^{q_1}(\alpha_1,b)}$$

for every $t > 0$ and $u_0 \in L^{q_1}(\alpha_1, b)$. Moreover, if I is a compact subset of $(0, \infty)$, the constants C_1 and C_2 can be chosen uniformly for all $c \in I$.

COROLLARY 2.10. *Let p, α_1, α_2, and b be as in Lemma 2.7. Assume that $1 < q_1 \leq q_2 < \infty$ and that $r \geq 2$. Then, there exists a $C > 0$ such that*

$$\|U_c(t)Q_c u_0\|_{L^{q_2}(\alpha_2,b)} \leq Ct^{-3/4}(1+t)^{-r+3/4}\|Q_c u_0\|_{L^{-1,q_1}(\alpha_1,b)}$$

for every $t > 0$ and $u_0 \in L^{-1,q_1}(\alpha_1, b)$. Moreover, if I is a compact subset of $(0, \infty)$, the constant C can be chosen uniformly for all $c \in I$.

The proofs of Lemma 2.7 and Corollaries 2.9 and 2.10 will be given in the appendix.

3. Separation of the motions. Let us represent the solution u to the Cauchy problem (1.1)–(1.3) as

$$u(x,t) = \varphi_{c(t)}(y) + v(y,t) \tag{3.1}$$

with $y = x - \int_0^t c(s)ds + \gamma(t)$.

In order to distinguish the motion of the solitary wave part and that of the residual part, we impose the constraint that

$$v(y,t) \in \text{Range}\left(Q_{c(t)}(t)\right).$$

This requirement corresponds to

$$\langle v(\cdot,t),\, \eta_i(\cdot,\, c(t))\rangle = 0 \quad \text{for } i = 1,\, 2, \tag{3.2}$$

which can be satisfied by modulating $c(t)$ and $\gamma(t)$. In this section, we show that the decomposition exists locally in time, and we derive an evolution equation which arises from (3.2).

PROPOSITION 3.1. *Let $p \geq 1$, $\alpha > 3/4$, and $t_0 \geq 0$. Suppose $u_0 \in H^1 \cap L^2(\alpha, b)$. Then there exist positive numbers δ_0 and δ_1 such that, for any real γ_0, if the solution u to (1.1) satisfies*

$$\sup_{0\leq t\leq t_0} \|u(\cdot - \gamma_0, t) - \varphi_{c_0}(\cdot - c_0 t)\|_{L^2(\alpha,b)} < \delta_0,$$

there exists a unique function $(c(t), \gamma(t)) \in C([0,t_0];\mathbb{R}^2) \cap C^1((0,t_0);\mathbb{R}^2)$ satisfying

$$\sup_{0\leq t\leq t_0} (|c(t) - c_0| + |\gamma(t) - \gamma_0|) < \delta_1, \tag{3.3}$$

$$F_i[u,c,\gamma] := \left\langle u\left(\cdot + \int_0^t c(s)ds - \gamma(t), t\right) - \varphi_{c(t)}(\cdot),\, \eta_i(\cdot, c(t))\right\rangle = 0 \tag{3.4}$$

for $i = 1,\, 2$ and $0 \leq t \leq t_0$. The number δ_0 may be chosen as the decreasing function of t_0.

Before we prove the proposition, we will show the continuity of the solutions in $L^2(\alpha, b)$.

LEMMA 3.2. *Let $p \geq 1$, $\alpha > 3/4$, and $b > 0$. If $u_0 \in H^1 \cap L^2(\alpha, b)$, the solution u to (1.1)–(1.2) satisfies*

$$u \in C([0,\infty); L^2(\alpha, b)) \cap C([0,\infty); H^1(\mathbb{R})).$$

Furthermore, if $\alpha > 2$ and $u_0 \in H^1 \cap L^2(\alpha, b)$, it holds that

$$u \in C((0,\infty); W^{1,2}(\alpha, b)).$$

Proof. As one easily sees, there exist σ and λ_0 with $\lambda_0 > \sigma > 0$ such that $-\partial_x^3 - \sigma$ is dissipative on $L^2(\alpha, b)$ and λ_0 belongs to the resolvent set $\rho(-\partial_x^3)$. So, the operator $-\partial_x^3$ generates a C_0-semigroup $T_0(t)$ on $L^2(\alpha, b)$ (see [24, p. 119] and [26]).

By the variation of constants formula, the solution u to (1.1) satisfies

$$u(t) = T_0(t)u_0 + \int_0^t T_0(t-s)\partial_x f(u(s))ds, \tag{3.5}$$

where $T_0(t) = \exp(-t\partial_x^3)$. To prove the former part of the proposition, we will show that each term of (3.5) belongs to $C([0,\infty); L^2(\alpha, b))$. Since $T_0(t)$ is a C^0-semigroup defined on $L^2(\alpha, b)$, it holds that $T_0(t)u_0 \in C([0,\infty); L^2(\alpha, b))$.

Next, we show that the second term of (3.5) belongs to $C([0,\infty); L^2(\alpha, b))$. For each t, the operator $T_0(t)$ can be represented as the convolution with the function

$$S_t(x) = (3t)^{-1/3}\,\mathrm{Ai}(x(3t)^{-1/3}),$$

where

$$\mathrm{Ai}(x) = (2\pi)^{-1}\int \exp(i\xi^3/3 + ix\xi)d\xi.$$

Let α_1 be a number with $\alpha_1 > \alpha + 1/4$. Since the Airy function $\mathrm{Ai}(x)$ satisfies

$$|\mathrm{Ai}^{(i)}(x)| \leq C_i(1+x_-)^{(2i-1)/4}e^{-Cx_+^{3/2}} \quad \text{for } i = 0, 1 \tag{3.6}$$

(see [17, p. 213]) and

$$|h_{\alpha,b}(x)h_{\alpha_1,b}(y)^{-1}| \leq \begin{cases} Ce^{b(x-y)} & \text{if } x \geq y, \\ C\langle x-y\rangle^{\alpha-\alpha_1} & \text{if } x \leq y \end{cases} \tag{3.7}$$

follows, then

$$\big\|\|h_{\alpha,b}(x)S_t(x-y)h_{\alpha_1,b}(y)^{-1}\|_{L^2(\mathbb{R}_x)}\big\|_{L^\infty(\mathbb{R}_y)} \leq Ct^{-1/3} \quad \text{for } t \in [0,T].$$

Combining this with Minkowski's inequality, we have

$$\|T_0(t)v\|_{L^2(\alpha,b)} \leq C_T t^{-1/3}\|v\|_{L^1(\alpha_1,b)} \quad \text{for } t \in [0,T], \tag{3.8}$$

where $\alpha_1 > \alpha + 1/4$, $T > 0$ is arbitrary, and C_T is a positive number depending on T. Since the solution u to (1.1)–(1.2) with $u_0 \in H^1 \cap L^2(\alpha_1, b)$ satisfies

$$u \in C(\mathbb{R}; H^1) \cap L^\infty_{loc}([0,\infty); L^2(\alpha, b)),$$
$$u_x \in L^2_{loc}([0,\infty); L^2(\alpha - 1/2, b))$$

(see [15], [20], and the references therein), we have

$$f(u)_x \in L^q_{loc}([0,\infty); L^1(\alpha_*, b)) \quad \text{for } \alpha_* = \alpha + (2\alpha - 1)/q \text{ and } 2 \le q \le \infty. \tag{3.9}$$

Combining (3.8) and (3.9) with $q = 2$ and the fact that $T_0(t)$ is the C^0-semigroup on $L^2(\alpha, b)$, we have

$$\begin{aligned}
&\left\| \int_0^{t+h} T_0(t+h-s)\partial_x f(u(s))ds - \int_0^t T_0(t-s)\partial_x f(u(s))ds \right\|_{L^2(\alpha,b)} \\
&\le \left\| \int_t^{t+h} T_0(t+h-s)\partial_x f(u(s))ds \right\|_{L^2(\alpha,b)} \\
&\quad + \left\| (T_0(h)-1) \int_0^t T_0(t-s)\partial_x f(u(s))ds \right\|_{L^2(\alpha,b)} = o(1) \quad \text{as } h \downarrow 0,
\end{aligned}$$

which implies the right continuity of the second term in $L^2(\alpha, b)$. The left continuity can be shown exactly in the same way.

Now, we prove the latter part of the proposition. Using (3.6) and (3.7), one can see that

$$\|\partial_x T_0(t)v\|_{L^2(\alpha,b)} \le Ct^{-3/4}\|v\|_{L^1(\alpha_1,b)} \quad \text{for } t \in [0,T]$$

with $0 < T < \infty$ if $\alpha_1 > \alpha + 3/4$. Combining this with (3.5) and (3.9) with $q > 4$, we can prove that $u(t) \in C((0,\infty); W^{1,2}(\alpha, b))$ in the same way. Thus we complete the proof. □

Proof of Proposition 3.1. The proof follows the line of Proposition 5.1 in [28]. Let $F[u, c, \gamma] = (F_1[u, c, \gamma], F_2[u, c, \gamma])$ be the functional defined by (3.4) that maps $C([0, t_0); L^{-3,2}(\alpha, b)) \times C([0, t_0); \mathbb{R}^2)$ to $C([0, t_0); \mathbb{R}^2)$. We remark that $F \in C^1$ and that

$$u \in C^1([0,\infty); L^{-3,2}(\alpha, b))$$

follows from Lemma 3.2 and (1.1).

Put $U_0 = (\varphi_{c_0}(\cdot - c_0 t), c_0, \gamma_0)$. Then it follows that $F[U_0] = 0$ and

$$D_{(c,\gamma)}F[U_0] = \begin{pmatrix} B & -1 \\ -1 & 0 \end{pmatrix},$$

where B is a functional given by $B[c](t) = \int_0^t c(s)ds$. Applying the implicit function theorem, we see that there exists a unique function $(c(t), \gamma(t))$ satisfying (3.3) and (3.4) and that $u \mapsto (\gamma(\cdot), c(\cdot))$ is a C^1-mapping from $C([0, t_0]; L^{-3,2}(\alpha, b))$ to $C([0, t_0]; \mathbb{R}^2)$. Since $u \in C^1([0,\infty); L^{-3,2}(\alpha, b))$, we conclude that $(c(t), \gamma(t)) \in C^1([0, t_0); \mathbb{R}^2)$. Thus we have proved Proposition 3.1. □

Let δ_0 be a positive number. Then, there is an $\varepsilon > 0$ such that if $\|v_0\|_{1,0} + \|v_0\|_{L^2(\alpha_2,b)} < \varepsilon$, we have $u \in C([0, t_1]; L^2(\alpha_2, b))$ with

$$\sup_{0 \le t \le t_1} \|u(\cdot - \gamma_0, t) - \varphi_{c_0}(\cdot - c_0 t)\|_{L^2(\alpha_2,b)} \le \delta_0$$

for some $t_1 > 0$. If δ_0 is a sufficiently small number, Proposition 3.1 implies the existence of decomposition (3.1)–(3.2) on the time interval $[0, t_1]$.

Next, we will derive the evolution equations of (v, c, γ), which are valid for $0 < t < t_1$. Substituting (3.1) into (1.1), we have

$$\partial_t v - A_{[c(t)]}v + \dot{\gamma}(t)\partial_y v + \partial_y N(t, v) + (\dot{c}(t)\partial_c + \dot{\gamma}(t)\partial_y)\varphi_{c(t)} = 0, \tag{3.10}$$

where

$$N(t, v) = f(\varphi_{c(t)} + v) - f(\varphi_{c(t)}) - f'(\varphi_{c(t)})v.$$

Let $\psi(t)$ be an arbitrary smooth function with $\psi(0) = \psi(t_1) = 0$, and let $\chi_j(x)$ $(j \in \mathbb{N})$ be nonnegative smooth functions satisfying

$$\chi_j(x) = \begin{cases} 0 & \text{for } x \geq j + 1, \\ 1 & \text{for } x \leq j. \end{cases}$$

Then

$$-\int_0^{t_1} \langle v(t),\, \chi_j \eta_i(\cdot, c(t))\rangle \psi' dt = \int_0^{t_1} \{\langle v_t,\, \chi_j \eta_i(\cdot, c(t))\rangle + \dot{c}\langle v,\, \chi_j \partial_c \eta_i(\cdot, c(t))\rangle\}\, \psi(t) dt. \tag{3.11}$$

Substituting (3.10) into the integrand of the right-hand side of (3.11), integrating the resulting equation by parts, sending $j \to \infty$, and using (2.4) and (3.2), we have

$$\mathcal{A}(t)\begin{pmatrix} \dot{\gamma} \\ \dot{c} \end{pmatrix} = -\begin{pmatrix} \langle N(t, v), \partial_y \eta_1(\cdot, c(t))\rangle \\ \langle N(t, v), \partial_y \eta_2(\cdot, c(t))\rangle \end{pmatrix}, \tag{3.12}$$

where

$$\mathcal{A}(t) = \begin{pmatrix} 1 - \langle v,\, \partial_y \eta_1(\cdot, c(t))\rangle & -\langle v,\, \partial_c \eta_1(\cdot, c(t))\rangle \\ -\langle v,\, \partial_y \eta_2(\cdot, c(t))\rangle & 1 - \langle v,\, \partial_c \eta_2(\cdot, c(t))\rangle \end{pmatrix}. \tag{3.13}$$

Thus we obtain a triplet of equations (3.10) and (3.12), which describes the motion of the solitary wave part and the residual part. We remark that the system (3.10) and (3.12) with $\langle v(0), \eta_i\rangle = 0$ $(i = 1, 2)$ is valid and equivalent to (1.1)–(1.3) as long as the decomposition (3.1) with (3.2) persists.

4. A priori estimates and asymptotic stability. In this section, we aim to prove Theorem 2.2. Let t_1 be a positive number such that the decomposition (3.1) with (3.2) persists for $0 < t < t_1$. In order to prove the theorem, we first derive from (3.10) and (3.12) a priori estimates of (v, c, γ) on $[0, t_1]$, which do not depend on t_1. The estimates will ensure the persistence of the decomposition (3.1) with (3.2) for all the time and imply that the residual part v locally decays to 0 as $t \to \infty$.

Let $c_1 = c(t_1)$ and $\gamma_1 = \gamma(t_1) - \int_0^{t_1}(c(s) - c_1)ds$. Using the change of variables

$$\tilde{v}(z, t) = v(y, t), \quad z = x - c_1 t + \gamma_1, \tag{4.1}$$

we can rewrite (3.10) into

$$\partial_t \tilde{v} - A_{[c_1]}\tilde{v} + \partial_z \widetilde{N}(t, \tilde{v}) + (\dot{c}\partial_c + \dot{\gamma}\partial_z)\varphi_{c(t)}(y) = 0, \tag{4.2}$$

where

$$\begin{aligned} \widetilde{N}(t, \tilde{v}) &= f(\varphi_{c(t)}(y) + \tilde{v}) - f(\varphi_{c(t)}(y)) - f'(\varphi_{c_1}(z))\tilde{v} \\ &= \int_0^1 \{f'(\varphi_{c(t)}(y) + \theta\tilde{v}) - f'(\varphi_{c_1}(z))\} d\theta \tilde{v}. \end{aligned}$$

To obtain the local energy decay estimate of $\tilde{v}$, we make use of the spectral properties of the linearized operator around $\varphi_{c_1}(z)$. Let us decompose $\tilde{v}$ into the contribution of the generalized eigenfunctions of $A_{[c_1]}$ and the other part. Let $g = P_{c_1}\tilde{v}$ and $h = Q_{c_1}\tilde{v}$. Then by (3.2) and (4.2),

$$h(t) = U_{c_1}(t)h(0) - \int_0^t U_{c_1}(t-s)Q_{c_1}\partial_z \widetilde{N}(s,\tilde{v})ds$$
$$- \int_0^t U_{c_1}(t-s)Q_{c_1}(\dot{c}\partial_c + \dot{\gamma}\partial_z)\varphi_{c(s)}(y)ds, \tag{4.3}$$

$$\mathcal{B}(t)\begin{pmatrix}\kappa_1\\ \kappa_2\end{pmatrix} + \begin{pmatrix}\int h(z,t)\eta_1(y,c(t))dz\\ \int h(z,t)\eta_2(y,c(t))dz\end{pmatrix} = 0, \tag{4.4}$$

and $\kappa_i(t) = \langle g(\cdot), \eta_i(\cdot, c_1)\rangle$ for $i = 1, 2$ and

$$\mathcal{B}(t) = \begin{pmatrix}\int \xi_1(z,c_1)\eta_1(y,c(t))dz & \int \xi_2(z,c_1)\eta_1(y,c(t))dz\\ \int \xi_1(z,c_1)\eta_2(y,c(t))dz & \int \xi_2(z,c_1)\eta_2(y,c(t))dz\end{pmatrix}.$$

From the orthogonality conditions (2.5), we see that $\mathcal{B}(t)$ satisfies

$$\mathcal{B}(t) = \begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix} + O(|c(t)-c_1| + |y(t)-z(t)|). \tag{4.5}$$

Now, let us introduce several functions:

$$M_0(t) = \sup_{0\le\tau\le t}\langle\tau\rangle^{2r-2}|y(\tau)-z(\tau)|, \qquad M_1(t) = \sup_{0\le\tau\le t}\langle\tau\rangle^{2r-1}|c(\tau)-c_1|,$$
$$M_2(t) = \sup_{0\le\tau\le t}\langle\tau\rangle^{r}\left(|\kappa_1(\tau)| + |\kappa_2(\tau)|\right), \qquad M_3(t) = \sup_{0\le\tau\le t}\|\tilde{v}(\cdot,\tau)\|_{1,0},$$
$$M_4(t) = \sup_{0\le\tau\le t}\langle\tau\rangle^{r}\|h(\cdot,\tau)\|_{L^2(\alpha_2,b)},$$
$$M_5(t) = \sup_{0\le\tau\le t}\tau^{3/4+\eta}\langle\tau\rangle^{r-3/4-\eta}\|\partial_z h(\cdot,\tau)\|_{L^2(\alpha_2,b)},$$

where η is a sufficiently small positive number.

Our strategy to prove Theorem 2.2 is to estimate the above system of functions and derive the closed system of inequalities that does not depend on t_1. Hereafter, we denote by $C(M)$ various functions of M_i $(0 \le i \le 5)$ which are bounded in some neighborhood of the origin.

PROPOSITION 4.1. *Let p, α_1, α_2, b, and r be positive numbers as in Theorem 2.2. Assume that $v_0 \in H^1 \cap L^2(\alpha_1, b)$. Then, for $t \in [0, t_1]$,*

$$M_0 + M_1 \le C(M)(M_2^2 + M_4^2), \tag{4.6}$$
$$M_2 \le C(M)M_4(M_0 + M_1), \tag{4.7}$$
$$M_3 \le C(\|v_0\|_{1,0} + \|v_0\|_{L^2(\alpha_2,b)}) + C(M)(M_2 + M_4), \tag{4.8}$$
$$\begin{aligned} M_4 + M_5 \le{}& \|\tilde{v}(0)\|_{L^2(\alpha_1,b)} + \delta(M_2 + M_4)\\ &+ C(M)(M_2 + M_4)(M_2 + M_3 + M_5)\\ &+ C(M)\{(M_0 + M_1)(M_2 + M_4 + M_5) + M_2^2 + M_4^2\}, \end{aligned} \tag{4.9}$$

where C is a positive constant and $\delta = \delta(M_0, M_1, M_2, M_3) \to 0$ as $\sum_{0\le i\le 3} M_i \to 0$. Moreover, the constants C, $C(M)$, and δ do not depend on t_1.

Proof. To begin with, we estimate $M_0(t)$ and $M_1(t)$. By (3.13), we have

$$\mathcal{A}(t) = I + O(\|v\|).$$

Since

$$\|v\|_{L^2_y(\alpha_2,b)} \le C(\|h\|_{L^2_z(\alpha_2,b)} + |\kappa_1| + |\kappa_2|)e^{bM_0},$$

it follows from (3.12) that

$$|\dot\gamma(\tau)| + |\dot c(\tau)| \le C(M)\|v\|^2_{L^2_y(\alpha_2,b)} \le C(M)(M_2^2 + M_4^2)\langle\tau\rangle^{-2r}. \tag{4.10}$$

So, we see that (4.6) follows from (4.10) and the definition of M_0 and M_1.

Using the orthogonality conditions $\langle h(\cdot,\tau), \eta_i(\cdot, c_1)\rangle = 0$ $(i = 1, 2)$, we have

$$\begin{aligned}\left|\int h(z,\tau)\eta_i(y,c(\tau))dz\right| &= \left|\int h(z,\tau)(\eta_i(y,c(\tau)) - \eta_i(z,c_1))dz\right| \\ &\le C(M)M_4(M_0+M_1)(1+\tau)^{-r}.\end{aligned}$$

Combining this with (4.4) and (4.5), we obtain (4.7).

We now turn to the estimate of $M_4(t)$ and $M_5(t)$. Let $0 < \beta \le 2\alpha_2$, and let χ_+ and χ_- be the characteristic functions of the intervals $[0,\infty)$ and $(-\infty, 0]$. Noting that

$$\begin{aligned}|\varphi^{(i)}_{c(\tau)}(y) - \varphi^{(i)}_{c_1}(z)| &\le C(M)(|y-z| + |c(\tau)-c_1|)(e^{-c(\tau)|y|} + e^{-c_1|z|}) \\ &\le C(M)(M_0+M_1)(e^{-c(\tau)|y|} + e^{-c_1|z|})\end{aligned}$$

for $i = 0$, 1 and that

$$\begin{aligned}\|\tilde v\tilde v_z\|_{L^1(\beta,b)} &\le C\|\chi_+\tilde v\|_{L^2(\alpha_2,b)}\|\chi_+\tilde v_z\|_{L^2(\alpha_2,b)} + C\|\chi_-\tilde v\|_{L^2(\alpha_2,b)}\|\chi_-\tilde v_z\| \\ &\le Cs^{-3/4-\eta}\langle s\rangle^{-r+3/4+\eta}(M_2+M_4)(M_2+M_3+M_5),\end{aligned}$$

we compute

$$\begin{aligned}&\|\partial_z \widetilde N(s,\tilde v)\|_{L^1(\beta,b)} \\ &\quad\le \left\|\int_0^1 \{f'(\varphi_{c(s)}(y)+\theta\tilde v) - f'(\varphi_{c_1}(z))\}d\theta\tilde v_z\right\|_{L^1(\beta,b)} \\ &\qquad + \left\|\int_0^1 f''(\varphi_{c(s)}(y)+\theta\tilde v)(\varphi'_{c(s)}(y) - \varphi'_{c_1}(z) + \theta\tilde v_z)d\theta\tilde v\right\|_{L^1(\beta,b)} \\ &\qquad + \left\|\int_0^1 \{f''(\varphi_{c(s)}(y)+\theta\tilde v) - f''(\varphi_{c_1}(z))\}\varphi'_{c_1}(z)d\theta\tilde v\right\|_{L^1(\beta,b)} \\ &\quad\le C(M)\{\|\tilde v\tilde v_z\|_{L^1(\beta,b)} + (|c(s)-c_1| + |y(s)-z(s)|)\|\tilde v\|_{W^{1,2}(\alpha_2,b)}\} + \delta\|\tilde v\|_{L^2(\alpha_2,b)} \\ &\quad\le s^{-3/4-\eta}\langle s\rangle^{-r+3/4+\eta}C(M)(M_2+M_4)(M_2+M_3+M_5) \\ &\qquad + s^{-3/4-\eta}\langle s\rangle^{-r+3/4+\eta}C(M)(M_0+M_1)(M_2+M_4+M_5) + \langle s\rangle^{-r}\delta(M_2+M_4),\end{aligned} \tag{4.11}$$

where $\delta = \sup_{0\le\theta\le1}|f''(\varphi_{c(s)}(y)+\theta\tilde v) - f''(\varphi_{c_1}(z))|$. We see that $\delta = 0$ if $f(u) = u^2$ and that δ is a positive number that tends to 0 as $\sum_{0\le i\le 3} M_i \to 0$ if $p > 1$.

Let $\alpha_1 = \alpha_2 + r + 1$. Applying Corollary 2.9 to (4.3) and using (4.10) and (4.11), we have

$$\begin{aligned}
&\|h\|_{L^2(\alpha_2,b)}\\
&\le C\langle t\rangle^{-r}\|Q_{c_1}\tilde{v}(0)\|_{L^2(\alpha_1,b)}\\
&\quad + C\int_0^t \langle t-s\rangle^{-r+1/4}|t-s|^{-1/4}\|Q_{c_1}\partial_z\widetilde{N}(s,\tilde{v})\|_{L^1(\alpha_1-1/2,b)}\\
&\quad + C\int_0^t \langle t-s\rangle^{-r}\|Q_{c_1}(\dot{c}\partial_c + \dot{\gamma}\partial_z)\varphi_{c(s)}(y)\|_{L^2(\alpha_1,b)}\\
&\le C\langle t\rangle^{-r}\|Q_{c_1}\tilde{v}(0)\|_{L^2(\alpha_1,b)} + \langle t\rangle^{-r}(M_2+M_4)\{\delta + C(M)(M_2+M_3+M_5)\}\\
&\quad + \langle t\rangle^{-r}C(M)\{(M_0+M_1)(M_2+M_4+M_5) + (M_2^2+M_4^2)\}
\end{aligned}$$

and

$$\begin{aligned}
&\|\partial_z h\|_{L^2(\alpha_2,b)}\\
&\le Ct^{-3/4}\langle t\rangle^{-r+3/4}\|Q_{c_1}\tilde{v}(0)\|_{L^2(\alpha_1,b)}\\
&\quad + C\int_0^t \langle t-s\rangle^{-r+3/4}|t-s|^{-3/4}\|Q_{c_1}\partial_z\widetilde{N}(s,\tilde{v})\|_{L^1(\alpha_1-1/2,b)}\\
&\quad + C\int_0^t \langle t-s\rangle^{-r+3/4}|t-s|^{-3/4}\|Q_{c_1}(\dot{c}\partial_c + \dot{\gamma}\partial_z)\varphi_{c(s)}(y)\|_{L^2(\alpha_1,b)}\\
&\le Ct^{-3/4}\langle t\rangle^{-r+3/4}\|Q_{c_1}\tilde{v}(0)\|_{L^2(\alpha_1,b)}\\
&\quad + t^{-1/2-\eta}\langle t\rangle^{-r+1/2+\eta}(M_2+M_4)\{\delta + C(M)(M_2+M_3+M_5)\}\\
&\quad + C(M)t^{-1/2-\eta}\langle t\rangle^{-r+1/2+\eta}(M_0+M_1)(M_2+M_4+M_5)\\
&\quad + C(M)\langle t\rangle^{-r}(M_2^2+M_4^2).
\end{aligned}$$

Thus we have (4.9). Furthermore, we get $\lim_{t\downarrow 0} M_5(t) = 0$.

Finally, we estimate the H^1-norm of the residual part. As in [28, p. 335], the H^1-estimate can be obtained by using the Lyapunov stability analysis and the local energy decay of v. Let E be the functional defined by

$$E[u] = \int \left(\frac{1}{2}u_x^2 + \frac{c_0}{2}u^2 - F(u)\right) dx,$$

where $F' = f$ with $F(0) = 0$. Put $k = u(x,t) - \varphi_{c_0}(y) = v(y,t) + \varphi_{c(t)}(y) - \varphi_{c_0}(y)$. Since φ_{c_0} is a critical point of the functional E,

$$E[u] - E[\varphi_{c_0}] = \frac{1}{2}\int \left(k_y^2 + c_0k^2 - 2f'(\varphi_{c_0})k^2\right) dy + O(\|k\|_{1,0}^3). \tag{4.12}$$

Note that $\delta E := E[u] - E[\varphi_{c_0}]$ is constant in time and that

$$|\delta E| = |E[\varphi_{c_0} + v_0] - E[\varphi_{c_0}]| \le C\|v_0\|_{1,0}^2.$$

As can be easily seen,

$$|\|k\|_{1,0} - \|v\|_{1,0}| \le C|c(t) - c_0|,$$

$$\int f'(\varphi_{c_0})k^2 dy \le C(\|v\|_{L^2(\alpha_2,b)} + |c(t) - c_0|)^2.$$

Combining these inequalities with (4.12), we have

$$\|v\|_{1,0}^2 \le C(\delta E + |c(t) - c_0|^2 + \|v\|_{L^2(\alpha_2,b)}^2). \tag{4.13}$$

By (4.10),

$$|c(t) - c_0| + |\gamma(t) - \gamma_0| \le |c(0) - c_0| + |\gamma(0) - \gamma_0| + C(M)(M_2^2 + M_4^2). \tag{4.14}$$

Applying Proposition 3.1 with $t_1 = 0$, we see that the map $v_0 \mapsto (c(0), \gamma(0))$ is C^1 in a small neighborhood of the origin. In particular, the map is locally Lipschitz continuous and

$$|c(0) - c_0| + |\gamma(0) - \gamma_0| \le C\|v_0\|_{L^2(\alpha_2,b)}. \tag{4.15}$$

Combining (4.13)–(4.15), we have (4.8). □

COROLLARY 4.2. *Let p, α_1, α_2, and b be positive numbers as in Theorem* 2.2. *Then, there exist $\varepsilon_0 > 0$ and $C > 0$ such that*

$$\|v_0\|_{1,0} + \|v_0\|_{L^2(\alpha_1,b)} = \varepsilon < \varepsilon_0$$

implies

$$\sum_{i=0}^{5} M_i(t) \le C\varepsilon \quad \textit{for every } t \ge 0.$$

Proof. In view of (4.15) and $\tilde{v}(z,0) = v_0(x) + \varphi_{c_0}(x + \gamma_0) - \varphi_{c(0)}(x + \gamma(0))$, we have

$$\|\tilde{v}(0)\|_{L^2_z(\alpha_1,b)} \le C(M)\|v_0\|_{L^2(\alpha_1,b)}. \tag{4.16}$$

By Proposition 4.1 and (4.16),

$$\sum_{i=0}^{5} M_i(t) \le C(M)\varepsilon + \delta(M_2(t) + M_4(t)) + W(M_0(t), \dots, M_5(t)), \tag{4.17}$$

where $W = O(\sum_{i=0}^{5} M_i^2)$ and $\delta = o(1)$ as $\sum_{i=0}^{5} M_i \to 0$. So, if ε is sufficiently small, the solution of (4.17) can belong either to a small neighborhood of 0 or to a domain whose distance from 0 is bounded from below uniformly with respect to ε. Since $M_i(t) \in C([0,t_1];\mathbb{R})$ $(0 \le i \le 5)$ follows Proposition 3.1, Lemma 3.2, and the fact that $\lim_{t\downarrow 0} M_5(t) = 0$, there exists a $C > 0$ such that

$$\sum_{i=0}^{5} M_i(t) \le C\varepsilon \quad \text{for } t \in [0,t_1]. \tag{4.18}$$

Let T be the supremum of the set of all positive numbers t_1, for which the solution $u(x,t)$ has a decomposition (3.1)–(3.2) for $t \in [0,t_1]$. The proof of Corollary 4.2 will be complete if we obtain $T=\infty$. Suppose that $T<\infty$ and put $c_*=c(T-t_0/2)$, $\gamma_*=-\int_0^{T-t_0/2}c(s)ds+\gamma(T-t_0/2)$, and $\tilde{u}(x,t)=u(x-\gamma_*,t+T-t_0/2)$, where t_0 is a positive number to be fixed later. By (4.6), (4.9), (4.15), and (4.18), we obtain

$$\begin{aligned}\|\tilde{u}(x,0)-\varphi_{c_0}(x)\|_{L^2(\alpha_2,b)} &\le \|v(x,T-t_0/2)\|_{L^2(\alpha_2,b)}+\|\varphi_{c_*}-\varphi_{c_0}\|_{L^2(\alpha_2,b)}\\ &\le C\varepsilon,\end{aligned}$$

where C is a positive number which does not depend on ε and t_0. Therefore, if ε_0 is sufficiently small, there exists a $t_0>0$ such that $\tilde{u}(x,t)$ satisfies the hypothesis of Proposition 3.1. (Note that δ_0 in Proposition 3.1 can be chosen as a decreasing function of t_0.) From $\tilde{u}$, we get $(\tilde{\gamma}(t),\tilde{c}(t))$ that satisfies $(\tilde{\gamma}(0),\tilde{c}(0))=(\gamma_*,c_*)$. Then the extension can be defined by $(\gamma(t),c(t))=(\tilde{\gamma}(t-T+t_0/2)-\gamma_*+\gamma(T-t_0/2),\tilde{c}(t+T-T_{\tilde{0}}/2))$ for $T-T_0/2\le t\le T+t_0/2$, which contradicts the definition of T. Thus we have $T=\infty$. □

Now, we are in position to prove Theorem 2.2.

Proof of Theorem 2.2. Corollary 4.2 and (4.10) imply that there exists a $C>0$ such that

$$|\dot{\gamma}(t)|+|\dot{c}(t)|\le C\varepsilon(1+t)^{-2r}\quad\text{for every } t\ge 0.$$

Hence,

$$c_+=\lim_{t\to\infty}c(t)\quad\text{and}\quad\gamma_+=\lim_{t\to\infty}\left(\gamma(t)-\int_0^t(c(s)-c_+)ds\right)$$

exist and satisfy $|c(t)-c_+|\le C\varepsilon(1+t)^{-2r+1}$ and $|y(t)-z(t)|\le C\varepsilon(1+t)^{-2r+2}$, where $z(t)=x-c_+t+\gamma_+$. Using Corollary 4.2 and the above estimates, we have

$$\begin{aligned}&\|u(\cdot+c_+t-\gamma_+,t)-\varphi_{c_+}(\cdot)\|_{W^{1,2}(\alpha_2,b)}\\ \le&\|\tilde{v}(x,t)\|_{W^{1,2}(\alpha_2,b)}+\|\varphi_{c(t)}(\cdot+y(t)-z(t))-\varphi_{c_+}\|_{W^{1,2}(\alpha_2,b)}\\ \le&C\varepsilon t^{-r}\end{aligned}$$

and

$$\|u(\cdot+c_+t-\gamma_+,t)-\varphi_{c_+}(\cdot)\|_{1,0}\le\|\tilde{v}(x,t)\|_{1,0}+\|\varphi_{c(t)}(\cdot+y(t)-z(t))-\varphi_{c_+}\|_{1,0}\le C\varepsilon.$$

In view of (4.10), (4.15), and Corollary 4.2, we have (2.7). Thus we complete the proof of Theorem 2.2. □

Finally, we will show the boundedness of

$$\begin{aligned}M_6&=\sup_{\tau\ge0}\langle\tau\rangle^{r-1/4}\tau^{1/4}\|h(\cdot,\tau)\|_{L^q(\alpha_3,b)},\\ M_7&=\sup_{\tau\ge0}\tau^{3/4}\langle\tau\rangle^{r-3/4}\|h_z(\cdot,\tau)\|_{L^q(\alpha_3,b)}\end{aligned}$$

for $2\le q\le\infty$, which shall be used in the next section.

LEMMA 4.3. *Let p, α_1, α_2, r, and b be positive numbers as in Theorem* 2.2. *Assume that $2\le q\le\infty$ and $1<\alpha_3\le r+1-1/q$. Then, there exist $C>0$ and $\varepsilon_0>0$ such that*

$$\|v_0\|_{1,0}+\|v_0\|_{L^2(\alpha_1,b)}=\varepsilon<\varepsilon_0$$

implies

$$M_6 + M_7 \le C\varepsilon.$$

Proof. Applying Corollary 2.9 to (4.3), we have

$$\begin{aligned}\|h\|_{L^q(\alpha_3,b)} &\le C\langle t\rangle^{-r+1/4}t^{-1/4}\|Q_{c_+}\tilde{v}(0)\|_{L^2(\alpha_1,b)}\\ &\quad + C\int_0^t \langle t-s\rangle^{-r+1/3}|t-s|^{-1/3}\|Q_{c_+}\partial_z\widetilde{N}(s,\tilde{v})\|_{L^1(\alpha_3+r+1/q,b)}\\ &\quad + C\int_0^t \langle t-s\rangle^{-r+1/4}|t-s|^{-1/4}\|Q_{c_1}(\dot{c}\partial_c+\dot{\gamma}\partial_z)\varphi_{c(s)}(y)\|_{L^2(\alpha_1,b)}\end{aligned}$$

and

$$\begin{aligned}\|h_z\|_{L^q(\alpha_3,b)} &\le C\langle t\rangle^{-r+3/4}t^{-3/4}\|Q_{c_+}\tilde{v}(0)\|_{L^2(\alpha_1,b)}\\ &\quad + C\int_0^t \langle t-s\rangle^{-r+3/4}|t-s|^{-3/4}\|Q_{c_+}\partial_z\widetilde{N}(s,\tilde{v})\|_{L^1(\alpha_3+r+1/q,b)}\\ &\quad + C\int_0^t \langle t-s\rangle^{-r+3/4}|t-s|^{-3/4}\|Q_{c_1}(\dot{c}\partial_c+\dot{\gamma}\partial_z)\varphi_{c(s)}(y)\|_{L^2(\alpha_1,b)}.\end{aligned}$$

Substituting (4.10) and (4.11) into the above inequalities and applying Corollary 4.2, we have $M_6 + M_7 \le C\varepsilon$. Thus we have proved Lemma 4.3. □

5. Scattering. In this section, we will prove Theorem 2.4. Let

$$u(x,t) = \varphi_{c_+}(z) + w(x,t) \quad \text{and} \quad z = x - c_+t + \gamma_+.$$

Then

$$w_t + w_{xxx} + f(\varphi_{c_+}+w)_x - f(\varphi_{c_+})_x = 0 \quad \text{for } x\in\mathbb{R} \text{ and } t>0. \tag{5.1}$$

The ingredient to prove the existence of scattering states is to show

$$\sup_{x\in\mathbb{R}}|w(x,t)| \le C(1+t)^{-1/3}. \tag{5.2}$$

Our strategy to obtain (5.2) is to combine the method due to Hayashi and Naumkin [16] with the local energy decay estimate obtained in section 4.

Let L and I be the operators defined by

$$L\phi = \partial_t\phi + \partial_x^3\phi, \quad I\phi = x\phi + 3t\int_{-\infty}^x \partial_t\phi dx', \quad J\phi = (x - 3t\partial_x^2)\phi \tag{5.3}$$

for $\phi\in C_0^\infty(\mathbb{R}^2)$. Note that

$$[L,J]=0, \quad [L,I] = 3\int_{-\infty}^x L\phi dx', \quad [I,\partial_x]\phi = [J,\partial_x]\phi = -\phi. \tag{5.4}$$

To prove (5.2), we need the following lemma.

LEMMA 5.1 (see Hayashi and Naumkin [16]). *Let*

$$N[u](t) := \|u(t)\|_{1,0} + \|D_x^\alpha Ju(t)\| + \|D_xJu(t)\| + |\overline{u}(t)|, \quad \overline{u}(t) = \int_{\mathbb{R}} u(x,t)dx.$$

If $q \in (4,\infty]$ and $u(x,t)$ is a smooth function with $\sup_{t\geq 0} N[u](t) < \infty$ for an $\alpha \in (0,1/2)$, then there exists a $C>0$ such that

$$\|u\|_{L^q} \leq C(1+t)^{-1/3+1/(3q)} N[u](t),$$
$$\|uu_x\|_{L^\infty} \leq Ct^{-2/3}(1+t)^{-1/3} N[u]^2(t)$$

for every $t \geq 0$.

Now, we will show the boundedness of $N[w]$.

PROPOSITION 5.2. *Let p, α be as in Theorem 2.4, and let β be a number with $\beta = 1/2-\gamma$ and $0<\gamma<\min\{1/2,(p-2)/3\}$. Then, there exist positive numbers ε_0 and C such that, for every $0<\varepsilon<\varepsilon_0$ and $t\geq 0$,*

$$\|(1+x_+)^\alpha v_0\| + \|v_0\|_{1,1} = \varepsilon < \varepsilon_0$$

implies

$$N[w](t) = \|w(t)\|_{1,0} + \|D^\beta Jw(t)\| + \|\partial_x Jw(t)\| + |\overline{w}(t)| \leq C\varepsilon. \tag{5.5}$$

Proof. To begin with, we remark that the solution u to (1.1)–(1.2) with $p>2$ and $u_0 \in H^{1,1}$ locally exists in time and satisfies $N[u](t)<\infty$ (see [16] and the references therein).

Now, we estimate each term of $N[w]$. Noting that

$$w(x,t) = \tilde{v}(z,t) + \varphi_{c(t)}(y) - \varphi_{c_+}(z), \tag{5.6}$$

and using Corollary 4.2 and (5.1), we have

$$\begin{aligned} \|w\|_{1,0} &\leq \|\tilde{v}(z,t)\|_{1,0} + \|\varphi_{c(t)}(y) - \varphi_{c_+}(z)\|_{1,0} \\ &\leq C(\|v_0\|_{1,0} + \|v_0\|_{L^2(\alpha_1,b)}) \end{aligned} \tag{5.7}$$

and

$$\begin{aligned} |\overline{w}(t)| = |\overline{w}(0)| &\leq C\|w(0)\|_{1,1} \\ &\leq C(\|v_0\|_{1,1} + \|v_0\|_{L^2(\alpha_1,b)}). \end{aligned} \tag{5.8}$$

By (5.1), (5.3), and (5.4),

$$\begin{aligned} LIw &= ILw + 3\int_{-\infty}^x Lw dx' \\ &= (x\partial_x + 3t\partial_t + 3)\{f(\varphi_{c+}) - f(\varphi_{c+}+w)\} \\ &= f'(\varphi_{c_+})I\varphi_{c_+,x} - f'(\varphi_{c_+}+w)I(\varphi_{c_+}+w)_x \\ &\quad + 3(f(\varphi_{c+}) - f(\varphi_{c+}+w)). \end{aligned} \tag{5.9}$$

Applying D^β to (5.9), taking an inner product of the resulting equation and $D^\beta Iw$ in L^2, and using (5.4), we have

$$\frac{d}{dt}\|D^\beta Iw\|^2 = -2(K_1+K_2+K_3+K_4+K_5), \tag{5.10}$$

where

$$
\begin{aligned}
K_1 &= (D^\beta Iw, D^\beta\{(f'(\varphi_{c_+}+w)-f'(\varphi_{c_+}))(I\varphi_{c_+})_x\}),\\
K_2 &= (D^\beta Iw, D^\beta\{f'(w)(Iw)_x\}),\\
K_3 &= (D^\beta Iw, D^\beta\{(f'(\varphi_{c_+}+w)-f'(w))(Iw)_x\}),\\
K_4 &= (D^\beta Iw, D^\beta\{g(\varphi_{c_+}+w)-g(\varphi_{c_+})-g(w)\}),\\
K_5 &= (D^\beta Iw, D^\beta g(w)),
\end{aligned}
$$

where $g(v)=3f(v)-f'(v)v$.

Multiplying (5.9) by $(Iw)_{xx}$ and integrating the resulting equation by parts, we have

$$
\frac{d}{dt}\|(Iw)_x\|^2 = -2(K_6+K_7+K_8+K_9), \tag{5.11}
$$

where

$$
\begin{aligned}
K_6 &= ((Iw)_x, \{(f'(\varphi_{c_+}+w)-f'(\varphi_{c_+}))(I\varphi_{c_+})_x\}_x),\\
K_7 &= \frac{1}{2}\int |(Iw)_x|^2 f''(w)w_x dx + \int (Iw)_x g'(w)w_x dx,\\
K_8 &= \frac{1}{2}\int |(Iw)_x|^2\{(f'(\varphi_{c_+}+w)-f'(w))\}_x dx,\\
K_9 &= ((Iw)_x, \{g(\varphi_{c_+}+w)-g(\varphi_{c_+})-g(w)\}_x).
\end{aligned}
$$

Using (5.6), $(I\varphi_{c_+})_x = (z-2c_+t-\gamma_+)\varphi'_{c_+}(z)+\varphi_{c_+}(z)$, and Corollary 4.2, we have

$$
\begin{aligned}
|K_1| &\le C\|D^\beta Iw\|\left\|\int_0^1 f''(\varphi_{c_+}+\theta w)d\theta w(I\varphi_{c_+})_x\right\|_{1,0}\\
&\le C(1+t)\|D^\beta Iw\|(\|\tilde v\|_{W^{1,2}(\alpha_1,b)}+|c(t)-c_+|+|y(t)-z(t)|)\\
&\le Ct^{-3/4}(1+t)^{-r+7/4}\|D^\beta Iw\|(M_0+M_1+M_4+M_5)\\
&\le C\varepsilon t^{-3/4}(1+t)^{-r+7/4}\|D^\beta Iw\|,
\end{aligned}
$$

$$
\begin{aligned}
|K_4| &\le C\|D^\beta Iw\|\left\|\int_0^1 g''(\theta_1\varphi_{c_+}+\theta_2 w)d\theta_1 d\theta_2\varphi_{c_+}w\right\|_{1,0}\\
&\le Ct^{-3/4}(1+t)^{-r+3/4}\|D^\beta Iw\|(M_0+M_1+M_4+M_5)\\
&\le C\varepsilon t^{-3/4}(1+t)^{-r+3/4}\|D^\beta Iw\|
\end{aligned}
$$

for $r\ge 2$. Here we used the fact that $\min\{r,2r-2\}=r$ for $r\ge 2$. Similarly, we have

$$
\begin{aligned}
|K_6| &\le C\varepsilon t^{-3/4}(1+t)^{-r+7/4}\|(Iw)_x\|,\\
|K_9| &\le C\varepsilon t^{-3/4}(1+t)^{-r+3/4}\|(Iw)_x\|.
\end{aligned}
$$

By Lemma 2.3 of [16], we have

$$\begin{aligned}|K_2| &\le C\|D^\beta Iw\|(\|D^\beta Iw\| + \|(Iw)_x\|)(\|w\|_\infty^{p-2}\|ww_x\|_\infty + \|w\|_\infty^{p-2-2\gamma}\|w\|^{2\gamma}\|ww_x\|_\infty \\ &\qquad + \|w\|_\infty^{p-2+2\gamma}\|ww_x\|_\infty^{1-\gamma}) \\ &\le C(N[w])^p t^{-2/3}\langle t\rangle^{-(p-1-2\gamma)/3}\|D^\beta Iw\|(\|D^\beta Iw\| + \|(Iw)_x\|), \\ |K_5| &\le C\|D^\beta Iw\|\|w\|_{2p}^p(\|ww_x\|_\infty^{1/2} + \|w\|_\infty^{3\gamma}\|ww_x\|_\infty^{(1-3\gamma)/2}) \\ &\le C(N[w])^{p+1}t^{-1/3}\langle t\rangle^{-p/3+\gamma/2}\|D^\beta Iw\|, \\ |K_7| &\le C\|w\|_\infty^{p-2}\|ww_x\|_\infty\|(Iw)_x\|(\|(Iw)_x\| + \|w\|) \\ &\le C(N[w])^p t^{-2/3}\langle t\rangle^{-(p-1)/3}(N[w] + \|(Iw)_x\|)\|(Iw)_x\|.\end{aligned}$$

To estimate K_3 and K_8, we will show the boundedness of

$$M_8(t) = \sup_{0\le\tau\le t} \tau^{-3/4}\langle\tau\rangle^{-r+7/4}\|(Iw)_x\|_{L^2(\alpha_4,b)},$$

where $\alpha_4 = \alpha_2 - 1$. Differentiating (5.9) by x, we have

$$L(Iw)_x + \partial_x(f'(\varphi_{c_+})(Iw)_x) = -\partial_x(I_1 + I_2), \tag{5.12}$$

where

$$\begin{aligned}I_1 &= \{f'(\varphi_{c_+} + w) - f'(\varphi_{c_+})\}(Iw)_x, \\ I_2 &= g(\varphi_{c_+} + w) - g(\varphi_{c_+}) + \{f'(\varphi_{c_+} + w) - f'(\varphi_{c_+})\}(I\varphi_{c_+})_x.\end{aligned}$$

Making use of the change of variables $z = x - c_+t + \gamma_+$ and the variation of constants formula, we can rewrite (5.12) into

$$\psi(t) = U_{c_+}(t)\psi(0) - \int_0^t U_{c_+}(t-s)\partial_z(I_1 + I_2)ds,$$

where $\psi(z,t) = (Iw)_x(x,t)$. Applying Corollary 2.10 to the equation above, we have

$$\begin{aligned}&\|Q_{c_+}\psi(t)\|_{L^2(\alpha_4,b)} \\ &\le \|U_{c_+}(t)Q_{c_+}\psi(0)\|_{L^2(\alpha_4,b)} + \int_0^t \|U_{c_+}(t-s)Q_{c_+}\partial_z(I_1+I_2)\|_{L^2(\alpha_4,b)}ds \\ &\le Ct^{-3/4}\langle t\rangle^{-r+3/4}\|\psi(0)\|_{L^{-1,2}(\alpha_1-1,b)} \\ &\quad + C\int_0^t |t-s|^{-3/4}\langle t-s\rangle^{-r+3/4}\|\partial_z(I_1+I_2)\|_{L^{-1,2}(\alpha_1-1,b)}.\end{aligned} \tag{5.13}$$

Clearly,

$$\|\psi(0)\|_{L^{-1,2}(\alpha_1-1,b)} \le \|xw(0)\|_{L^2(\alpha_1-1,b)} \le C\|(1+x_+)^{\alpha_1}w(0)\|. \tag{5.14}$$

Since $\|\partial_z I_2\|_{L^{-1,2}(\alpha_1-1,b)} \le C(1+s)\|w\|_{L^2(\alpha_2,b)}$ and

$$\begin{aligned}\|\partial_z I_1\|_{L^{-1,2}(\alpha_1-1,b)} &\le C\|(|\varphi_{c_+}| + |w|)^{p-1}w(Iw)_x\|_{L^2(\alpha_1-1,b)} \\ &\le C\|(Iw)_x\|\|w\|_{L^\infty(\alpha_3,b)}(1 + \|w\|_{L^\infty(\alpha_3,b)}\|w\|_\infty^{p-2})\end{aligned}$$

for $\alpha_3 = (\alpha_1 - 1)/2 = r + 1/4$, it follows from (5.13), (5.14), Corollary 4.2, and Lemma 4.3 that

$$\begin{aligned}\|Q_{c_+}\psi(t)\|_{L^2(\alpha_4,b)} &\le Ct^{-3/4}\langle t\rangle^{-r+3/4}\|(1+x_+)^{\alpha_1}w(0)\| \\ &\quad + C\left\{M_4 + \sup_{0\le s\le t}\|(Iw)_x\|(M_0+M_1+M_2+M_6)\right\} \\ &\quad \times \int_0^t |t-s|^{-3/4}\langle t-s\rangle^{-r+3/4}s^{-2/3}\langle s\rangle^{-r+5/3}ds \\ &\le C\varepsilon t^{-3/4}\langle t\rangle^{-r+7/4}\left(1+\sup_{0\le s\le t}\|(Iw)_x\|\right).\end{aligned} \tag{5.15}$$

On the other hand, by (2.6), (5.1), and (5.3),

$$\begin{aligned}\|P_{c_+}\psi(t)\|_{L^2(\alpha_4,b)} &\le \sum_{i=1}^{2}\|\langle (xw)_x + 3tw_t, \eta_i\rangle\xi_i\|_{L^2(\alpha_4,b)} \\ &\le C(1+t)\|w\|_{L^2(\alpha_2,b)}.\end{aligned} \tag{5.16}$$

Hence, it follows from (5.15) and (5.16) that

$$M_8(t) \le C\varepsilon\left(1+\sup_{0\le s\le t}\|(Iw)_x\|\right) \quad \text{for } t \ge 0.$$

Now, we turn to the estimate of K_3 and K_8. It follows that

$$\begin{aligned}|K_3| &\le (\|D^\beta Iw\| + \|(Iw)_x\|)\left\|\int_0^1 f''(\theta\varphi_{c_+}+w)d\theta\varphi_{c_+}(Iw)_x\right\| \\ &\le C(\|D^\beta Iw\| + \|(Iw)_x\|)\|\varphi_{c_+}(Iw)_x\| \\ &\le C\varepsilon t^{-3/4}\langle t\rangle^{-r+7/4}(\|D^\beta Iw\| + \|(Iw)_x\|)\left(1+\sup_{0\le s\le t}\|(Iw)_x\|\right).\end{aligned}$$

Furthermore, we have

$$\begin{aligned}|K_8| &\le C\int |(Iw)_x|^2|\left\{\int_0^1 f''(\theta\varphi_{c_+}+w)d\theta\varphi_{c_+}\right\}_x|dx \\ &\le C\|(Iw)_x\|(\|(Iw)_x\|_{L^2(\alpha_4,b)} + \|(Iw)_x\|\|w_x\|_{L^\infty(\alpha_3,b)}) \\ &\le Ct^{-3/4}\langle t\rangle^{-r+7/4}(M_8 + M_7\|(Iw)_x\|)\|(Iw)_x\| \\ &\le C\varepsilon t^{-3/4}\langle t\rangle^{-r+7/4}\left(1+\sup_{0\le s\le t}\|(Iw)_x\|\right)\|(Iw)_x\|.\end{aligned}$$

Substituting the above estimates of K_i $(6 \le i \le 9)$ into (5.11), we have

$$\begin{aligned}\frac{d}{dt}\|(Iw)_x\| \le& C\{\varepsilon t^{-3/4}\langle t\rangle^{-r+7/4} + (N[w])^{p+1}t^{-2/3}\langle t\rangle^{-(p-1)/3}\} \\ &+ C\{\varepsilon t^{-3/4}\langle t\rangle^{-r+7/4} + t^{-2/3}\langle t\rangle^{-(p-1)/3}(N[w])^p\}\sup_{0\le s\le t}\|(Iw)_x\|.\end{aligned}$$

Applying the Gronwall inequality, we have

$$\sup_{0\le s\le t}\|(Iw)_x\| \le C\sup_{0\le s\le t}\left\{(\varepsilon + (N[w])^p)\exp\left(\varepsilon + (N[w])^p\right)\right\}.$$

Substituting the estimates of K_i $(1 \le i \le 5)$ and that of $\|(Iw)_x\|$ into (5.10) and applying the Gronwall inequality, we have

$$\sup_{0\le s\le t} \|D^\beta Iw\| \le C\left(\sup_{0\le s\le t} N[w], \varepsilon\right) \sup_{0\le s\le t} \{\varepsilon + (N[w])^p\},$$

where C is bounded in some neighborhood of $(0,0)$. Combining these with (5.3) and Corollary 4.2, we have

(5.17)
$$\begin{aligned}\|D^\beta Jw\| + \|(Jw)_x\| &\le \|D^\beta Iw\| + \|(Iw)_x\| + 3t\|D^\beta\{f(\varphi_{c_+}+w) - f(\varphi_{c_+})\}\| \\ &\quad + 3t\|\{f(\varphi_{c_+}+w) - f(\varphi_{c_+})\}_x\| \\ &\le C\left(\sup_{0\le s\le t} N[w], \varepsilon\right) \sup_{0\le s\le t} (\varepsilon + (N[w])^p).\end{aligned}$$

Hence, if ε is sufficiently small, the inequality (5.5) follows from (5.7), (5.8), and (5.17). Thus we complete the proof. □

We are now in position to prove our main result.

Proof of Theorem 2.4. It follows from (5.1), Corollary 4.2, and Proposition 5.2 that

$$\begin{aligned}\|T(-t)w(t) - T(-s)w(s)\| &\le \int_s^t \|\{f(\varphi_{c_+}+w) - f(\varphi_{c_+})\}_x\| d\tau \\ &\le C\int_s^t (\|w\|_{W^{1,2}(\alpha_1,b)} + \|w\|_\infty^{p-2}\|ww_x\|_\infty\|w\|) d\tau \\ &\le C\varepsilon s^{-(p-2)/3}\end{aligned}$$

for $1 \le s \le t$. Therefore, there exists a $V \in L^2$ satisfying (2.10). This completes the proof of Theorem 2.4. □

Appendix A. In this section, we give the proofs of Lemma 2.7 and Corollaries 2.9 and 2.10, which is a generalization of Theorem 1.1 in [24]. The main difference between the weighted L^p-L^q estimate stated in section 2 and that of Miller [24] is that the indices p and q need not be equal in our case.

To start, let us introduce several notations. For Banach spaces X and Y, we denote the space of linear continuous operator by $\mathcal{L}(X,Y)$ and the space of linear compact operator by $\mathcal{L}_C(X,Y)$. We abbreviate $\mathcal{L}(X,X)$ and $\mathcal{L}_C(X,X)$ as $\mathcal{L}(X)$ and $\mathcal{L}_C(X)$, respectively. Let $A_{0,c} = -\partial_x^3 + c\partial_x$ and $T_c = A_{[c]} - A_{0,c}$. We define the resolvent operator of $A_{0,c}$ and $A_{[c]}$ as $R_0(\lambda;c) = (\lambda - A_{0,c})^{-1}$ and $R(\lambda;c) = (\lambda - A_{[c]})^{-1}$, respectively. For a linear subspace W of X, we denote by ${}^\perp W$ the subspace $\{g \in X^* \,|\, \langle g, u\rangle = 0 \text{ for every } u \in W\}$.

The proof of Lemma 2.7 basically follows the lines of Miller [24]. However, there is a mistake to be fixed in [24], although it is not so serious. In fact, she claims that $(1 - R_0(0;c)T_c)^{-1} \in \mathcal{L}(L^p(\alpha,b), Q_c L^p(\alpha,b))$ in Proposition 3.7 and Lemma 3.6 of [24]. But the Fredholm alternative theorem tells us that

$$(1 - R_0(0;c)T_c)u = A_{0,c}\varphi_c$$

does not have any solution in $L^p(\alpha,b)$ because $\mathrm{Ker}(1 - R_0(0;c)T_c)^* = \{\alpha A_{0,c}\varphi_c \,|\, \alpha \in \mathbb{C}\}$. For the sake of the reader's convenience, we shall give the complete proof of the resolvent expansion of $R(\lambda;c)$ around $\lambda = 0$ following the lines of Murata [25].

The first step is an examination of the free resolvent $R_0(\lambda;c)$ on $\Sigma := \{\lambda \,|\, \mathrm{Re}\,\lambda > 0\}$.

LEMMA A.1.

(i) *Let $m = 0,\ 1,\ 2,\ 3,\ k,\ r \in \mathbb{N} \cup \{0\}$, and $1 \le q_1 \le q_2 \le \infty$. Let $\alpha_1 \ge \alpha_2 + r + 1 + 1/q_2 - 1/q_1$, $\alpha_2 > 0$, and $b > 0$ be a small number. Then $R_0^{(i)}(\lambda;c) \in C\left(\overline{\Sigma} \times (0,\infty); \mathcal{L}(W^{k,q_1}(\alpha_1,b), W^{k+3,q_2}(\alpha_2,b))\right)$ for $0 \le i \le r$ and there exists a $C > 0$ such that*

$$\|R_0^{(i)}(\lambda;c)\|_{\mathcal{L}(W^{k,q_1}(\alpha_1,b);W^{k+m,q_2}(\alpha_2,b))} \le C\langle\lambda\rangle^{-2(i+1)/3+m/3} \tag{A.1}$$

for $\lambda \in \overline{\Sigma}$. For every compact subset I of $(0,\infty)$, the constant C in (A.1) *can be chosen uniformly for all $c \in I$.*

(ii) *Let r, m, α_1, α_2, and b be as in* (i). *Let $k \in \mathbb{Z}$ and $1 < q_1 \le q_2 < \infty$. Then there exists a $C > 0$ such that*

$$\|R_0^{(i)}(\lambda;c)\|_{\mathcal{L}(L^{k,q_1}(\alpha_1,b);L^{k+m,q_2}(\alpha_2,b))} \le C\langle\lambda\rangle^{-2(i+1)/3+m/3}. \tag{A.2}$$

For every compact subset I of $(0,\infty)$, the constant C in (A.2) *can be chosen uniformly for all $c \in I$.*

Proof. The equation

$$\mu^3 - c\mu + \lambda = 0$$

has roots $\mu_i(\lambda;c)$ $(i = 1,2,3)$ which satisfy

$$\mathrm{Re}\,\mu_1(\lambda;c) < 0 < \mathrm{Re}\,\mu_2(\lambda;c) \le \mathrm{Re}\,\mu_3(\lambda;c)$$

for $\lambda \in \Sigma$ and

$$\mathrm{Re}\,\mu_1(\lambda;c) < 0 = \mathrm{Re}\,\mu_2(\lambda;c) < \mathrm{Re}\,\mu_3(\lambda;c)$$

for $\lambda \in \partial\Sigma$. The roots $\mu_i(\lambda;c)$ $(i = 1,2,3)$ are continuous in λ and c. Moreover, they satisfy

$$\mu_1(\lambda;c) = -\sqrt{c} + O(\lambda), \quad \mu_2(\lambda;c) = \lambda/c + O(\lambda^3), \quad \mu_3(\lambda;c) = \sqrt{c} + O(\lambda)$$

around $\lambda = 0$ and

$$\mu_i(\lambda;c) = (-\lambda)^{1/3} + O(|\lambda|^{-1/3}) \tag{A.3}$$

uniformly as $\lambda \to \infty$ in Σ (see [27], [28]). So if I is a compact subset of $(0,\infty)$, we have $\mu_* = \sup_{c\in I} \sup_{\lambda\in\Sigma} \mathrm{Re}\,\mu_1(\lambda;c) < 0$.

Let $K_0(x,\lambda;c)$ be the kernel of $R_0(\lambda;c)$. Then, for $\lambda \in \overline{\Sigma}$ with $\lambda \ne 2(c/3)^{3/2}$ (i.e., $\mu_2 \ne \mu_3$), the kernel of $R_0(\lambda;c)$ is given by

$$K_0(x,\lambda;c) = \begin{cases} a_1(\lambda;c)e^{\mu_1(\lambda;c)x} & \text{for } x \ge 0, \\ -a_2(\lambda;c)e^{\mu_2(\lambda;c)x} - a_3(\lambda;c)e^{\mu_3(\lambda;c)x} & \text{for } x \le 0, \end{cases} \tag{A.4}$$

where $a_i(\lambda;c) = \prod_{j\ne i}(\mu_i(\lambda;c) - \mu_j(\lambda;c))^{-1}$. Put

$$\widetilde{K}_{r,m}(x,y,\lambda;c) = |h_{\alpha_2,b}(x)\partial_\lambda^r \partial_x^m K_0(x-y,\lambda;c) h_{\alpha_1,b}(y)^{-1}|.$$

From the definition of $h_{\alpha,b}$, (A.4), and the fact that

$$\frac{d^r}{d\lambda^r}(\mu_i^m a_i) = O(\langle\lambda\rangle^{(m-2r-2)/3}),$$

it follows that

$$\widetilde{K}_{r,m}(x,y,\lambda;c) \le \begin{cases} C\langle\lambda\rangle^{(m-2r-2)/3}\langle x\rangle^{\alpha_2}\langle y\rangle^{-\alpha_1+r} & \text{if } 0 \le x \le y, \\ C\langle\lambda\rangle^{(m-2r-2)/3}e^{bx/2}\langle y\rangle^{-\alpha_1+r} & \text{if } x \le 0 \text{ and } y \ge 0, \\ C\langle\lambda\rangle^{(m-2r-2)/3}e^{b(x-y)}\langle x-y\rangle^{r} & \text{if } x \le y \le 0, \\ C\langle\lambda\rangle^{(m-2r-2)/3}\langle x-y\rangle^{r}e^{(\mu_*+b)(x-y)} & \text{if } x \ge y, \end{cases}$$

where C is a positive number that does not depend on λ, x, and y. Moreover, since μ_i and its derivatives are continuous in λ and c and (A.3) holds uniformly with respect to $c \in I$, the constant C can be chosen uniformly for some neighborhood of $c \in (0,\infty)$. Hence, we have

$$\sup_{x\in\mathbb{R}}\|\widetilde{K}_{r,m}(x,\cdot,\lambda;c)\|_{L^{q_3}} + \sup_{y\in\mathbb{R}}\|\widetilde{K}_{r,m}(\cdot,y,\lambda;c)\|_{L^{q_3}} \le C\langle\lambda\rangle^{\{m-2(r+1)\}/3} \tag{A.5}$$

for $m = 0,1,2$, where $1/q_3 = 1 + 1/q_2 - 1/q_1$ and C is a positive number that can be chosen uniformly in some neighborhood of c. Furthermore, we have

$$\lim_{(c_1,\lambda_1)\to(c,\lambda)} |||\widetilde{K}_{r,m}(\cdot,\cdot,\lambda_1;c_1) - \widetilde{K}_{r,m}(\cdot,\cdot,\lambda;c)||| \to 0,$$

where $|||f(\cdot,\cdot)||| = \sup_{x\in\mathbb{R}}\|f(x,\cdot)\|_{L^{q_3}} + \sup_{y\in\mathbb{R}}\|f(\cdot,y)\|_{L^{q_3}}$. Using Young's inequality, we see that (A.5) implies (A.1) and that $\partial_x^m R_0(i)(\lambda;c)$ $(0 \le i \le r)$ is continuous in $\lambda \in \overline{\Sigma}$ and $c > 0$ if $k \in \mathbb{N}\cup\{0\}$ and $m = 0,1,2$. The case where $m = 3$ follows by using $\partial_x^3 R_0(\lambda;c) = 1 - (\lambda - c\partial_x)R_0(\lambda;c)$.

Next, we show (A.2). Using

$$(R_0^{(r)}(\lambda;c))^* u = \int_{\mathbb{R}} \partial_\lambda^r K_0(y-x,\bar{\lambda};c)u(y)dy,$$

we have

$$\|(R_0^{(r)}(\lambda;c))^*\|_{\mathcal{L}(L^{-k-m,q_2'}(-\alpha_2,-b);L^{-k,q_1'}(-\alpha_1,-b))} \le C\langle\lambda\rangle^{\{m-2(r+1)\}/3}$$

for $k \le -m$, where $1/q_1 + 1/q_1' = 1/q_2 + 1/q_2' = 1$ and C is a positive number. Hence, we have (A.2) with $k \le -m$ by using the standard duality argument. For $k \ge 0$, (A.2) immediately follows from (A.1) and the equivalence of $L^{k_i,q_i}(\alpha_i,b)$ and $W^{k_i,q_i}(\alpha_i,b)$ for $i = 0,1$. Combining these, we obtain (A.2).

Looking at the above proof, we see that $R_0^{(i)}(\lambda;c)$ is continuous in λ and c and the constant C in (A.1) and (A.2) can be chosen uniformly for some neighborhood of $c \in (0,\infty)$. □

The next step is to estimate the resolvent operator $R(\lambda;c)$ by using the resolvent identity

$$R(\lambda;c) = (1 - R_0(\lambda;c)T_c)^{-1}R_0(\lambda;c) \tag{A.6}$$

around $\lambda = 0$. The proof of the next lemma follows the lines of Murata [25].

LEMMA A.2. *Assume that $p \neq 4$. Let b, r, and m be as in Lemma* A.1, *and let $\alpha_1 \geq \alpha_2 + r + 1 + 1/q_2 - 1/q_1$, $\alpha_2 > 1 - 1/q_2$. Let I be a compact subset I of $(0, \infty)$.*

(i) *Suppose that $k \in \mathbb{N} \cup \{0\}$ and that $1 \leq q_1 \leq q_2 < \infty$ and $0 \leq i \leq r$. Then, there exists a neighborhood U of 0 in $\mathbb{C}$ such that*

$$\sup_{c \in I} \sup_{\lambda \in U \cap \overline{\Sigma}} \|Q_c R^{(i)}(\lambda; c)\|_{\mathcal{L}(W^{k,q_1}(\alpha_1,b); W^{k+m,q_2}(\alpha_2,b))} < \infty.$$

(ii) *Suppose that k is an integer with $k \geq -m$ and that $1 < q_1 \leq q_2 < \infty$. Then, there exists a neighborhood U of 0 in $\mathbb{C}$ such that*

$$\sup_{c \in I} \sup_{\lambda \in U \cap \overline{\Sigma}} \|Q_c R^{(i)}(\lambda; c)\|_{\mathcal{L}(L^{k,q_1}(\alpha_1,b); L^{k+m,q_2}(\alpha_2,b))} < \infty.$$

Proof. Let $X = W^{k,q_1}(\alpha_1, b)$ ($X = L^{k,q_1}(\alpha_1, b)$) and $Y = W^{k+m,q_2}(\alpha_2, b)$ ($Y = L^{k+m,q_2}(\alpha_2, b)$), respectively. Put $G_j = \frac{1}{j!} \frac{d^j}{d\lambda^j} R_0(0; c) = (-1)^j G_0^{j+1}$ for $j \in \mathbb{N} \cup \{0\}$. Because $f'(\varphi_c)$ is an exponentially decaying function and $\mu_i(\lambda; c)$ ($i = 0, 1, 2, 3$) are analytic in λ, it follows from (A.4) that

$$R_0(\lambda; c) T_c : \{\lambda \in \mathbb{C} \,|\, \operatorname{Re} \lambda > -\epsilon\} \to \mathcal{L}(Y)$$

is an analytic operator-valued function for an $\epsilon > 0$. By Lemma A.1, we see that $1 - R_0(\lambda; c) T_c$ is invertible in $\mathcal{L}(Y)$ for some $\lambda \in \Sigma$ with $|\lambda|$ sufficiently large. Hence, the analytic Fredholm theorem implies that $(1 - R_0(\lambda; c) T_c)^{-1}$ is a meromorphic function in $\mathcal{L}(Y)$ and has a Laurent series expansion

$$(1 - R_0(\lambda; c) T_c)^{-1}(\lambda) = \sum_{j=k}^{\infty} \lambda^j S_j$$

on $\{\lambda \in \mathbb{C} \,|\, 0 < |\lambda| < \delta\}$ for some $k \in \mathbb{Z}$ and $\delta > 0$. By Propositions 1.15 and 1.16 in [27], the singularity of $R(\lambda; c)$ is the same as the order of zero of the Evans function $D(\lambda)$. Hence, from (2.7) in [27],

$$\|R(\lambda; c)\|_{\mathcal{L}(X,Y)} = O(\lambda^{-2}) \quad \text{as } \lambda \to 0 \text{ in } \overline{\Sigma}. \tag{A.7}$$

In view of (A.7) and the identity

$$(1 + R(\lambda; c) T_c)(1 - R_0(\lambda; c) T_c) = 1,$$

we have $k \geq -2$. Thus we have

$$(1 - R_0(\lambda; c) T_c)^{-1} = S_{-2} \lambda^{-2} + S_{-1} \lambda^{-1} + r(\lambda; c), \tag{A.8}$$

where $r(\lambda; c)$ is an analytic function from $\{\lambda \in \mathbb{C} \,|\, \operatorname{Re} \lambda > -\epsilon\}$ to $\mathcal{L}(Y)$.

Since $(1 - R_0(\lambda; c) T_c)(1 - R_0(\lambda; c) T_c)^{-1} = (1 - R_0(\lambda; c) T_c)^{-1}(1 - R_0(\lambda; c) T_c) = 1$, it follows that, for $j = 0, -1, -2$,

$$(1 - G_0 T_c) S_j - G_1 T_c S_{j-1} - G_2 T_c S_{j-2} = \delta_j, \tag{A.9}$$

$$S_j (1 - G_0 T_c) - S_{j-1} G_1 T_c - S_{j-2} G_2 T_c = \delta_j, \tag{A.10}$$

where $S_j = 0$ for $j \leq -3$. By (2.3) and Proposition 2.1(ii),

$$\operatorname{Ker}(1 - G_0 T_c) = \operatorname{span}\{\partial_y \varphi_c\} \quad \text{and} \quad \operatorname{Ker}(1 - G_0 T_c)^* = \operatorname{span}\{A_{0,c} \varphi_c\}.$$

So (A.9) and (A.10) with $j=-2$ imply

$$\mathrm{Range} S_{-2} \subset \mathrm{Ker}(1-G_0T_c) = \mathrm{span}\{\partial_y\varphi_c\}, \tag{A.11}$$

$$\mathrm{Ker} S_{-2} \supset \mathrm{Range}(1-G_0T_c) = {}^{\perp}\mathrm{span}\{A_{0,c}\varphi_c\}. \tag{A.12}$$

Thus we have

$$S_{-2}v = \alpha\langle v, A_{0,c}\varphi_c\rangle\partial_y\varphi_c \tag{A.13}$$

for an $\alpha\in\mathbb{C}$. Applying $A_{0,c}$ to (A.9) with $j=-1$ and using $(1-G_0T_c)S_{-2}=0$, we have $A_{[c]}S_{-1}=S_{-2}$. Hence (A.13), (2.3), and Proposition 2.1(ii) yield

$$S_{-1}v = -\alpha\langle v, A_{0,c}\varphi_c\rangle\partial_c\varphi_c + \langle g, v\rangle\partial_y\varphi_c \tag{A.14}$$

for some $g\in Y^*$.

Since $Q_cS_{-2}=Q_cS_{-1}=0$, it follows from (A.8) that the operator $Q_c(1-R_0(\lambda;c)T_c)^{-1}$ is analytic in some neighborhood of $\lambda=0$.

Suppose that $c\in I$ and that $|\lambda|\le\varepsilon/4$. Then,

$$Q_c(1-R_0(\lambda;c)T_c)^{-1} = \frac{1}{2\pi i}\int_{|\zeta|=\frac{\varepsilon}{2}} \frac{Q_c(1-R_0(\zeta;c)T_c)^{-1}}{\zeta-\lambda}.$$

Since the integrand is continuous in c and λ and uniformly bounded on $\{\zeta\,|\,|\zeta|=\varepsilon/2\}$, we have $\sup_{c\in I}\sup_{|\lambda|\le\varepsilon/4}\|Q_c(1-R_0(\lambda;c)T_c)^{-1}\|_{\mathcal{L}(Y)}<\infty$. Combining the above with Lemma A.1 and (A.6), we have $\sup_{c\in I}\sup_{|\lambda|\le\varepsilon/4}\|Q_cR^{(r)}\|_{\mathcal{L}(X,Y)}<\infty$. Thus we have proved the lemma. □

LEMMA A.3. *Let p, m, r, α_1, α_2, and b be as in Lemma* A.2.

(i) *Let $1\le q_1\le q_2<\infty$ and $k\in\mathbb{N}\cup\{0\}$. Then, there exists a positive number C such that*

$$\|Q_cR^{(r)}(\lambda;c)\|_{\mathcal{L}(W^{k,q_1}(\alpha_1,b),W^{k+m,q_2}(\alpha_2,b))}\le C\langle\lambda\rangle^{-2(r+1)/3+m/3}$$

for every $\lambda\in\overline{\Sigma}$. Furthermore, if I is a compact subset of $(0,\infty)$, the constant C can be chosen uniformly for all $c\in I$.

(ii) *Let $1<q_1\le q_2<\infty$, and let k be an integer with $k\ge -m$. Then, there exists a positive number C such that*

$$\|Q_cR^{(r)}(\lambda;c)\|_{\mathcal{L}(L^{k,q_1}(\alpha_1,b),L^{k+m,q_2}(\alpha_2,b))}\le C\langle\lambda\rangle^{-2(r+1)/3+m/3}$$

for every $\lambda\in\overline{\Sigma}$. Furthermore, if I is a compact subset of $(0,\infty)$, the constant C can be chosen uniformly for all $c\in I$.

Proof. Let X and Y be as in the proof of Lemma A.2. First, we show that $(1-R_0(\lambda;c)T_c)^{-1}\in\mathcal{L}(Y)$ for every $\lambda\in\overline{\Sigma}\setminus\{0\}$. Let $\lambda\in\overline{\Sigma}$ be a point so that $1-R_0(\lambda;c)T_c$ is not invertible. Since $R_0(\lambda;c)T_c$ is a compact operator, it then follows that there exists a nontrivial solution $u\in Y$ to

$$u - R_0(\lambda;c)T_cu = 0. \tag{A.15}$$

It follows from Lemma A.1 and (A.15) that u is a classical solution to (2.1) and satisfies $u(x)=O((1+x)^{-\alpha_2})$ as $x\to\infty$ and $u(x)=O(e^{bx})$ as $x\to-\infty$. If $b>0$ is sufficiently small, [27, Proposition 1.6 and Theorem 3.6] tells us that the solution u actually decays exponentially as $|x|\to\infty$ and λ is also an eigenvalue of the operator

$A_{[c]}$ in L^2. Since 0 is the only eigenvalue of $A_{[c]}$ in L^2, it follows that $\lambda = 0$ and that $(1 - R_0(\lambda; c)T_c)^{-1} \in \mathcal{L}(Y)$ for $\lambda \in \overline{\Sigma} \setminus \{0\}$.

Now, let I be a compact subset of $(0, \infty)$. Then by Lemma A.1, there exists an $M > 0$ such that $\sup_{c\in I} \sup_{|\lambda|\le M} \|R_0(\lambda; c)T_c\|_{\mathcal{L}(Y)} \le 1/2$. For such an $M > 0$, we have

$$\sup_{c\in I} \sup_{|\lambda|\ge M} \|Q_c R^{(i)}(\lambda; c)T_c\|_{\mathcal{L}(X,Y)} < \infty \quad \text{for } 0 \le i \le r.$$

By Lemma A.1, the definitions of Q_c and T_c, and (A.6), $Q_c R^{(i)}(\lambda; c) \in \mathcal{L}(X, Y)$ $(0 \le i \le r)$ is continuous in $\lambda \in \overline{\Sigma} \setminus \{0\}$ and $c > 0$. The continuity of $Q_c R^{(i)}(\lambda; c)$ $(0 \le i \le r)$ and Lemma A.2 imply

$$\sup_{c\in I} \sup_{|\lambda|\le M} \|Q_c R^{(i)}(\lambda; c)T_c\|_{\mathcal{L}(X,Y)} < \infty \quad \text{for } 0 \le i \le r.$$

Thus we complete the proof. □

We are now in position to prove Lemma 2.7 and Corollaries 2.9 and 2.10 stated in section 2.

Proof of Lemma 2.7. The standard semigroup theory (see [26]) tells us that there exists a $\gamma > 0$ such that

$$e^{tA}u_0 = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{(\gamma+i\lambda)t} R(\gamma + i\lambda; c) d\lambda u_0$$

for $u_0 \in D(A^2) = H^6$. Let X and Y be as in the proof of Lemma A.2, and let $r > (m+1)/2$. Then Lemma A.3 implies that $\|R^{(r)}(\gamma + i\lambda)\|_{\mathcal{L}(X,Y)}$ decays uniformly at the rate $\langle\lambda\rangle^{-1-\varepsilon}$ as $\lambda \to \infty$, where $\varepsilon = (2r - m - 1)/3 > 0$. Let u, $v \in C_0^\infty(\mathbb{R})$. Since $R^{(r)}(\gamma + i\lambda) \in L^1(\mathbb{R}_\lambda; \mathcal{L}(X, Y))$ for every $\gamma > 0$, we may shift the contour of the integration and obtain

$$(e^{tA}Q_c u, v) = \frac{1}{2\pi(it)^r} \int_{-\infty}^{\infty} e^{it\lambda} (R^{(r)}(i\lambda + 0; c)Q_c u, v) d\lambda.$$

This implies that

$$\|U_c(t)Q_c\|_{\mathcal{L}(X,Y)} \le Ct^{-r}$$

for every $t > 0$. Thus we prove Lemma 2.7 for $r \in \mathbb{N}$ satisfying $r \ge [(m+1)/2] + 1$.

Let $\psi(x)$ be a smooth function with $\operatorname{supp}\psi = [1, 4]$, $\psi(x) > 0$ for $x \in (1, 4)$, and let

$$\psi_j(x) = \begin{cases} \psi(2^{-j}x) & \text{for } j \ge 0, \\ \psi\Big(\frac{3bx}{2\alpha \log 2} - \frac{3j}{2}\Big) & \text{for } j \le -1, \end{cases} \qquad \varphi_j(x) = \frac{\psi_j(x)}{\sum_{k=-\infty}^{\infty} \psi_k(x)}.$$

Then, $\varphi_j(x)$ $(j \in \mathbb{Z})$ are nonnegative functions satisfying the following:

(i) $$\operatorname{supp}\varphi_j = \begin{cases} [2^j, 2^{j+2}] & \text{for } j \ge 0, \\ [\alpha(j + 2/3) \log 2/b, \alpha(j + 8/3) \log 2/b] & \text{for } j \le -1, \end{cases}$$

(ii) $$\sum_{j=-\infty}^{\infty} \varphi_j(x) = 1,$$

(iii) $$\sup_{j\in\mathbb{Z}} \sup_{x\in\mathbb{R}} |\phi_j^{(i)}(x)| < \infty \quad \text{for every } i \in \mathbb{N} \cup \{0\}.$$

Put $\tilde{\varphi}_j = \sum_{k=-l}^{l} \varphi_{j+k}$. Let l be an integer such that $\varphi_j = \tilde{\varphi}_j \varphi_j$ for every $j \in Z$. We denote by A_j ($j \in \mathbb{Z}$) the Banach space $W^{k,p}(\alpha, b)$ equipped with the norm $\|\cdot\|_{A_j} = 2^{\alpha j}\|\cdot\|_{W^{k,p}}$. Set $Sf = \{\varphi_j(x)f(x)\}_{j=-\infty}^{\infty}$ and $R\{f_j\} = \sum_{j=-\infty}^{\infty} \tilde{\varphi}_j f_j$. Then by definition, $RS = E$ (the identity map), S is a bounded linear operator from $W^{k,p}(\alpha, b)$ to $\dot{l}_p(A_j)$, and R is a bounded linear operator from $\dot{l}_p(A_j)$ to $W^{k,p}(\alpha, b)$. Indeed,

$$\sum_{j\in\mathbb{Z}} \|\varphi_j f\|_{A_j}^p \le C \sum_{j\in\mathbb{Z}} \sum_{|\alpha|\le k} 2^{\alpha j p} \int_{\operatorname{supp}\varphi_j} |D^\alpha u|^p dx \le C\|u\|_{W^{k,p}(\alpha,b)}^p$$

and

$$\left\| \sum_{j\in\mathbb{Z}} \tilde{\varphi}_j f_j \right\|_{W^{k,p}(\alpha,b)}^p \le \sum_{j\in\mathbb{Z}} (6l+1)\|\tilde{\varphi}_j f_j\|_{W^{k,p}(\alpha,b)}^p \le C \sum_{j\in\mathbb{Z}} \|f_j\|_{A_j}^p.$$

Hence, R is a retraction. Combining the above with [2, Theorems 6.4.2 and 3.1.2] and the proof of [2, Theorem 5.6.1], we have Lemma 2.7. □

Proof of Corollaries 2.9 *and* 2.10. Let $T(t)$ denote the evolution operator corresponding to $-\partial_x^3 + c\partial_x$. Then it holds that

$$(T(t)f)(x) = \int_{\mathbb{R}} S_t(x - y + ct) f(y) dy \quad \text{for } t > 0. \tag{A.16}$$

Let

$$I^i(t, x, y) = |h_{\alpha_2,b}(x) D_x^i S_t(x - y + ct) h_{\alpha_1,b}(y)^{-1}|$$

and let $z = x - y$. By (3.6) and (3.7),

$$I^i(t,x,y) \le \begin{cases} C_1 t^{-(i+1)/3} \exp\left(bz - C_2 t^{-1/2}(z+ct)^{3/2}\right) & \text{if } z \ge 0, \\ C_1 t^{-(i+1)/3} \exp\left(-C_2 t^{-1/2}(z+ct)^{3/2}\right) \langle z\rangle^{\alpha_2-\alpha_1} & \text{if } -ct \le z \le 0, \\ C_1 t^{-(i+1)/3} \left(1 + t^{-1/3}|z+ct|\right)^{(2i-1)/4} \langle z\rangle^{\alpha_2-\alpha_1} & \text{if } z \le -ct, \end{cases}$$

where $i = 0, 1$ and C_i ($i = 1, 2$) are positive numbers. Let $1 \le q_1 \le q_2 \le \infty$. By the assumption, it holds that $q_3(\alpha_2 - \alpha_1) \le -1 - rq_3$, where $1/q_3 = 1 + 1/q_2 - 1/q_1$. Furthermore, if $x \ge y$ and b is sufficiently small, there exist positive numbers c_1 and c_2 such that

$$bz - C_2 t^{-1/2}(z+ct)^{3/2} \le -c_1 z^{3/2} t^{-1/2} - c_2 t.$$

Hence, there exists a $C > 0$ such that

$$\|I^0(t,x,y)\|_{L_y^\infty L_x^{q_3}} + \|I^0(t,x,y)\|_{L_x^\infty L_y^{q_3}} \le \begin{cases} C\langle t\rangle^{-r} t^{-1/4} \text{ if } 1 \le q_3 < 4, \\ C\langle t\rangle^{-r} t^{-1/3} \text{ if } q_3 \ge 4. \end{cases}$$

So applying Young's inequality to (A.16), we have

$$\|T(t)\|_{\mathcal{L}(L^{q_1}(\alpha_1,b), L^{q_2}(\alpha_2,b))} \le C\langle t\rangle^{-r} t^{-\theta} \quad \text{for } t > 0, \tag{A.17}$$

where $\theta = 1/3$ if $1/q_1 - 1/q_2 \le 3/4$ and $\theta = 1/4$ if $1/q_1 - 1/q_2 > 3/4$. Analogously, we have

$$\|I^1(t,x,y)\|_{L^\infty_y L^{q_3}_x} + \|I^1(t,x,y)\|_{L^\infty_x L^{q_3}_y} \le C\langle t\rangle^{-r+1/4} t^{-3/4}$$

for $t > 0$. Hence, it holds that

$$\|\partial_x T(t)\|_{\mathcal{L}(L^{q_1}(\alpha_1,b),L^{q_2}(\alpha_2,b))} \le C\langle t\rangle^{-r+1/4} t^{-3/4}. \tag{A.18}$$

Using that

$$\frac{d}{dt}\big(T(-t)U_c(t)Q_c\big) = T(-t)\partial_x\{f'(\varphi_c)U_c(t)Q_c u\}$$

for $u \in H^3$, and the fundamental theorem of the calculus, we compute

$$U_c(t)Q_c = T(t)Q_c - \int_0^t T(t-s)\partial_x\{f'(\varphi_c)U_c(s)Q_c\}ds. \tag{A.19}$$

Let

$$\begin{aligned} L_1(t) &= \sup_{0<s\le t}\left(s^\theta \|U_c(s)Q_c u_0\|_{L^{q_2}(\alpha_2,b)}\right), \\ L_2(t) &= \sup_{0<s\le t}\left(s^{3/4}\|\partial_x U_c(s)Q_c u_0\|_{L^{q_2}(\alpha_2,b)}\right). \end{aligned}$$

Combining (A.17)–(A.19), we have

$$\begin{aligned} &\|U_c(t)Q_c u_0\|_{L^{q_2}(\alpha_2,b)} \\ &\le \|T(t)Q_c u_0\|_{L^{q_2}(\alpha_2,b)} + \int_0^t \|\partial_x T(t-s)f'(\varphi_c)U_c(s)Q_c u_0\|_{L^{q_2}(\alpha_2,b)} \\ &\le Ct^{-\theta}\|u_0\|_{L^{q_1}(\alpha_1,b)} + CL_1(t)\int_0^t (t-s)^{-3/4}s^{-\theta}ds \\ &\le Ct^{-\theta}\|u_0\|_{L^{q_1}(\alpha_1,b)} + Ct^{1/4-\theta}L_1(t) \end{aligned}$$

and

$$\begin{aligned} &\|\partial_x U_c(t)Q_c u_0\|_{L^{q_2}(\alpha_2,b)} \\ &\le \|\partial_x T(t)Q_c u_0\|_{L^{q_2}(\alpha_2,b)} + \int_0^t \|\partial_x T(t-s)\partial_x\{f'(\varphi_c)U_c(s)Q_c u_0\}\|_{L^{q_2}(\alpha_2,b)} \\ &\le Ct^{-3/4}\|u_0\|_{L^{q_1}(\alpha_1,b)} + CL_2(t)\int_0^t (t-s)^{-3/4}s^{-3/4}ds \\ &\le Ct^{-3/4}\|u_0\|_{L^{q_1}(\alpha_1,b)} + Ct^{-1/2}L_2(t). \end{aligned}$$

Hence, there exist positive numbers t_1 and C such that

$$L_1(t_1) + L_2(t_1) \le C\|u_0\|_{L^{q_1}(\alpha_1,b)}. \tag{A.20}$$

Combining these with Lemma 2.7, we obtain Corollary 2.9 with $q_2 \ne \infty$.

Now, let $q_2 = \infty$. From Lemma 2.7 and (A.20), we have

$$\|U_c(t)Q_c u_0\|_{L^{q_1}(\alpha_1-r-1,b)} \le Ct^{-1/4}\langle t\rangle^{-r+1/4}\|u_0\|_{L^{q_1}(\alpha_1,b)}.$$

Using (A.18) and the above, we have

$$\begin{aligned}
&\|U_c(t)Q_cu_0\|_{L^\infty(\alpha_2,b)}\\
&\le \|T(t)Q_cu_0\|_{L^\infty(\alpha_2,b)} + \int_0^t \|\partial_x T(t-s)f'(\varphi_c)U_c(s)Q_cu_0\|_{L^\infty(\alpha_2,b)}\\
&\le Ct^{-\theta}\langle t\rangle^{-r}\|u_0\|_{L^{q_1}(\alpha_1,b)}\\
&\quad + C\int_0^t (t-s)^{-3/4}\langle t-s\rangle^{-r+1/4}\|f'(\varphi_c)U_c(s)Q_cu_0\|_{L^{q_1}(\alpha_1,b)}ds\\
&\le C\|u_0\|_{L^{q_1}(\alpha_1,b)}\left(\langle t\rangle^{-r}t^{-\theta} + \int_0^t (t-s)^{-3/4}\langle t-s\rangle^{-r+1/4}\langle s\rangle^{-r+1/4}s^{-1/4}ds\right)\\
&\le Ct^{-\theta}\langle t\rangle^{-r+\theta}\|u_0\|_{L^{q_1}(\alpha_1,b)}.
\end{aligned}$$

We can show

$$\|\partial_x U_c(t)Q_cu_0\|_{L^\infty(\alpha_2,b)} \le Ct^{-3/4}\langle t\rangle^{-r+3/4}\|u_0\|_{L^{q_1}(\alpha_1,b)}$$

in the same way.

Now, we turn to the proof of Corollary 2.10. Analogously to (A.18), we see that there exists a $C>0$ such that

$$\|T^*(t)u_0\|_{L^{1,q_1'}(-\alpha_1,-b)} \le Ct^{-3/4}\|u_0\|_{L^{q_2'}(-\alpha_2,-b)} \quad \text{for } t>0,$$

which is equivalent to

$$\|T(t)u_0\|_{L^{q_2}(\alpha_2,b)} \le Ct^{-3/4}\|u_0\|_{L^{-1,q_1}(\alpha_1,b)} \quad \text{for } t>0.$$

Using the above and (A.19), we see that there exist a $t_1>0$ and a $C>0$ such that

$$\sup_{0<s\le t_1} s^{3/4}\|U_c(s)Q_cu_0\|_{L^{q_2}(\alpha_2,b)} \le C\|u_0\|_{L^{-1,q_1}(\alpha_1,b)}$$

for every $u_0 \in L^{-1,q_1}(\alpha_1,b)$. On the other hand, in view of Lemma A.3 and the proof of Lemma 2.7, we get

$$\|U_c(t)Q_cu_0\|_{L^{q_2}(\alpha_2,b)} \le Ct^{-r}\|u_0\|_{L^{-1,q_1}(\alpha_1,b)} \quad \text{for } t>0.$$

Combining these, we obtain Corollary 2.10.

Looking at the proof, we see that constants C and C_i $(i=1,2)$, which appear in the right-hand side of the estimates, can be taken uniformly if the parameter c belongs to some compact subset of $(0,\infty)$. Thus we have completed the proof. □

Acknowledgments. The author would like to express his gratitude to Professor Hiroshi Matano for his encouragement and to Professor Yoshio Tsutsumi for stimulating discussions and valuable suggestions.

REFERENCES

[1] M. J. Ablowitz and H. Segur, *Solitons and the Inverse Scattering Transform*, Stud. Appl. Math. 4, SIAM, Philadelphia, PA, 1981.
[2] J. Bergh and J. Löfström, *Interpolation Spaces*, Springer-Verlag, Berlin, New York, 1976.
[3] T. B. Benjamin, *The stability of solitary waves*, Proc. Roy. Soc. London Ser. A, 328 (1972), pp. 153–183.

[4] J. L. Bona, *On the stability of solitary waves*, Proc. Roy. Soc. London Ser. A, 344 (1975), pp. 363–374.
[5] J. L. Bona, *On solitary waves and their role in the evolution of long waves*, in Applications of Nonlinear Analysis in the Physical Sciences, H. Amann, N. Bazley, and K. Kirchgässner, eds., Pitman, London, 1981, pp. 183–205.
[6] J. L. Bona, V. A. Dougalis, O. A. Karakashian, and W. R. McKinney, *The effect of dissipation on solutions of the generalized Korteweg-de Vries equation*, J. Comput. Appl. Math., 74 (1996), pp. 127–154.
[7] J. L. Bona, V. A. Dougalis, O. A. Karakashian, and W. R. McKinney, *Numerical simulation of singular solutions of the generalized Korteweg-de Vries equation*, in Mathematical Problems in the Theory of Water Waves Korteweg-de Vries Equation, Luminy, 1995, Contemp. Math. 200, AMS, Providence, RI, 1996, pp. 17–29.
[8] J. L. Bona, P. E. Souganidis, and W. A. Strauss, *Stability and instability of solitary waves of Korteweg-de Vries type*, Proc. Roy. Soc. London Ser. A, 411 (1987), pp. 395–412.
[9] J. L. Bona and A. Soyeur, *On the stability of solitary-wave solutions of model equations for long waves*, J. Nonlinear Sci., 4 (1994), pp. 449–470.
[10] V. S. Buslaev and G. S. Perelman, *Scattering for the nonlinear Schrödinger equation: States close to a soliton*, St. Petersburg Math. J., 4 (1993), pp. 1111–1142.
[11] F. M. Christ and M. I. Weinstein, *Dispersion of small amplitude solutions of the generalized Korteweg-de Vries equation*, J. Funct. Anal., 100 (1991), pp. 87–109.
[12] P. Deift and X. Zhou, *A steepest descent method for oscillatory Riemann-Hilbert problems. Asymptotics for the MKdV equation*, Ann. of Math., 137 (1993), pp. 295–368.
[13] C. S. Gardner, J. M. Greene, M. D. Kruskal, and R. M. Miura, *Method for solving the Korteweg-de Vries equation*, Phys. Rev. Lett., 19 (1967), pp. 1095–1097.
[14] C. S. Gardner, J. M. Greene, M. D. Kruskal, and R. M. Miura, *Korteweg-de Vries equation and generalizations. VI. Methods for exact solution*, Comm. Pure Appl. Math., 27 (1974), pp. 97–133.
[15] J. Ginibre, Y. Tsutsumi, and G. Velo, *Existence and uniqueness of solutions for the generalized Korteweg-de Vries equation*, Math. Z., 203 (1990), pp. 9–36.
[16] N. Hayashi and P. Naumkin, *Large time asymptotics of solutions to the generalized Korteweg-de Vries equation*, J. Funct. Anal., 159 (1998), pp. 110–136.
[17] L. Hörmonder, *The Analysis of Linear Partial Differential Operators* I, Springer-Verlag, Berlin, Heidelberg, 1983.
[18] A. Jensen and T. Kato, *Spectral properties of Schrödinger operators and time-decay of the wave functions*, Duke Math. J., 46 (1979), pp. 583–611.
[19] T. Kato, *On the Cauchy problem for the (generalized) Korteweg-de Vries equation*, Studies in Appl. Math., 8 (1983), pp. 93–128.
[20] C. Kenig, G. Ponce, and L. Vega, *Well-posedness and scattering results for the generalized Korteweg-de Vries equation via the contraction principle*, Comm. Pure Appl. Math., 46 (1993), pp. 527–620.
[21] S. Klainerman, *Long time behavior of solutions to nonlinear evolution equations*, Arch. Rational Mech. Anal., 78 (1982), pp. 73–89.
[22] S. Klainerman and G. Ponce, *Global small amplitude solutions to nonlinear evolution equations*, Comm. Pure. Appl. Math., 36 (1983), pp. 133–141.
[23] D. J. Korteweg and G. de Vries, *On the change of form of long waves advancing in a rectangular canal and on a new type of long stationary waves*, Philos. Mag. Ser. 5, 39 (1895), pp. 422–443.
[24] J. R. Miller, *Spectral properties and time decay for an Airy operator with potential*, J. Differential Equations, 141 (1997), pp. 102–121.
[25] M. Murata, *Asymptotic expansions in time for solutions of Schrödinger-type equations*, J. Funct. Anal., 49 (1982), pp. 10–56.
[26] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Appl. Math. Sci. 44, Springer-Verlag, New York, Berlin, Heidelberg, 1983.
[27] R. L. Pego and M. I. Weinstein, *Eigenvalues and instabilities of solitary waves*, Philos. Trans. Roy. Soc. London A, 340 (1992), pp. 47–94.
[28] R. L. Pego and M. I. Weinstein, *Asymptotic stability of solitary waves*, Comm. Math. Phys., 164 (1994), pp. 305–349.
[29] G. Ponce and L. Vega, *Nonlinear small data scattering for the generalized Korteweg-de Vries equation*, J. Funct. Anal., 90 (1990), pp. 445–457.
[30] M. A. Rammaha, *On the asymptotic behavior of solutions of generalized Korteweg-de Vries equations*, J. Math. Anal. Appl., 140 (1989), pp. 228–240.

[31] P. C. Schuur, *Asymptotic Analysis of Soliton Problems*, Lecture Notes in Math. 1232, Springer-Verlag, Berlin, Heidelberg, New York, 1986.
[32] J. Shatah, *Global existence of small solutions to nonlinear evolution equations*, J. Differential Equations, 46 (1982), pp. 409–425.
[33] A. Soffer and M. I. Weinstein, *Multichannel nonlinear scattering theory for nonintegrable equations*, Comm. Math. Phys., 133 (1990), pp. 119–146.
[34] A. Soffer and M. I. Weinstein, *Multichannel nonlinear scattering theory for nonintegrable equations* II: *The case of anisotropic potentials and data*, J. Differential Equations, 98 (1992), pp. 376–390.
[35] W. A. Strauss, *Dispersion of low-energy waves for two conservative equations*, Arch. Rational Mech. Anal., 55 (1974), pp. 86–92.
[36] W. A. Strauss, *Nonlinear scattering theory*, in Scattering Theory in Mathematical Physics, J. A. Lavita and J. P. Marchand, eds., Reidel Publishing, Dordrecht, Holland, 1974, pp. 53–78.
[37] M. I. Weinstein, *Lyapunov stability of ground states of nonlinear dispersive evolution equations*, Comm. Pure Appl. Math., 39 (1986), pp. 51–68.

SIAM J. MATH. ANAL.
Vol. 32, No. 5, pp. 1081–1093

OPTIMAL INEQUALITIES FOR GRADIENTS OF SOLUTIONS OF ELLIPTIC EQUATIONS OCCURRING IN TWO-PHASE HEAT CONDUCTORS*

ROBERT LIPTON†

Abstract. We consider solutions to divergence form partial differential equations that model steady state heat conduction in random two-phase composites. The coefficient representing the conductivity takes two scalar values. Optimal bounds on the L^2 norm of the gradient of the solution are found. The optimal upper bound is given in terms of the volume fraction occupied by each conducting phase. The optimal lower bound is independent of the volume fractions of the component conductors. The bounds follow from a Stieltjes integral representation for the L^2 norm of the gradient. Maximizing sequences of configurations are found using the corrector theory of homogenization.

Key words. homogenization, Stieltjes functions, spectral theorem, isoperimetric inequalities

AMS subject classifications. 35J, 35P, 74Q05

PII. S0036141000366625

1. Introduction.

Consider a bounded region Ω of R^N with a sufficiently regular boundary containing two isotropic conductors subjected to a constant applied temperature gradient $\mathbf{E}$ in R^N. Here we consider any dimension N greater than or equal to 2. The conductivities of the two materials are written as α and β, and the indicator function of the β phase is denoted by χ, where $\chi = 1$ inside the β phase and 0 otherwise. We suppose that the set occupied by the β phase is Lebesgue measurable and that $\beta > \alpha$. The local conductivity of the two-phase conductor is described by $a(\chi) = \alpha(1-\chi) + \beta\chi$. The temperature T inside the two-phase conductor is the solution of

$$-\text{div } (a(\chi)\nabla T) = 0 \tag{1.1}$$

subject to the boundary condition $T = \mathbf{E}\cdot\mathbf{x}$. Since the coefficient $a(\chi)$ is bounded and measurable, the equilibrium equation (1.1) is interpreted in the weak sense. Here we recall that the weak solution of (1.1) is defined to be the function T in $W^{1,2}(\Omega)$ that satisfies

$$\int_\Omega a(\chi)\nabla T\cdot\nabla\varphi\, dx = 0 \tag{1.2}$$

for all functions φ in $W_0^{1,2}(\Omega)$.

We suppose that the composite is random in that we specify only the volume fraction θ of the β phase and consider the ensemble of configurations that satisfy

*Received by the editors January 24, 2000; accepted for publication (in revised form) October 3, 2000; published electronically February 21, 2001. This research effort is sponsored by the NSF through grant DMS-9700638 and by the Air Force Office of Scientific Research, Air Force Materiel Command, United States Air Force, under grant F49620-99-1-0009. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the author and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Air Force Office of Scientific Research or the U.S. Government.

http://www.siam.org/journals/sima/32-5/36662.html

†Department of Mathematical Sciences, Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA 01609 (lipton@wpi.edu).

this isoperimetric constraint. The set of conductivities associated with this class is denoted by ad_θ and is written

$$ad_\theta = \left\{ a(\chi), \text{ where } \chi \text{ satisfies } \int_\Omega \chi \, dx = \theta \times meas(\Omega),\ 0 \le \theta \le 1 \right\}. \tag{1.3}$$

In this paper we address the problem of extremizing

$$\|\nabla T\|_2^2 \triangleq \int_\Omega |\nabla T|^2 \, dx \tag{1.4}$$

over the class ad_θ. We provide optimal upper and lower bounds on the quantity $\|\nabla T\|_2^2$. The upper bound depends explicitly upon the volume fraction occupied by the β phase. In order to state the bounds we set $\lambda = \frac{\beta}{\alpha}$ and introduce the function $f(z)$ defined by

$$f(z) = \frac{z}{\left(\frac{1}{1-\lambda} - z\right)^2} \tag{1.5}$$

and we give the following optimal inequality.

THEOREM 1.1 (optimal inequality for the L^2 norm of the gradient). *For any admissible conductivity $a(\chi)$ in ad_θ the associated temperature gradient ∇T satisfies*

$$meas(\Omega) \times |\mathbf{E}|^2 \le \|\nabla T\|_2^2 \le U(\theta, \mathbf{E}), \tag{1.6}$$

where $U(\theta, \mathbf{E})$ depends upon the contrast λ and is given by

$$U(\theta, \mathbf{E}) = meas(\Omega) \times (1 + \theta f(1-\theta)) \, |\mathbf{E}|^2 \quad \text{for } \lambda \le 2, \tag{1.7}$$

and for $\lambda \ge 2$

$$U(\theta, \mathbf{E}) = meas(\Omega) \times \begin{cases} (1 + \theta f(1/(\lambda-1))) \, |\mathbf{E}|^2 & \text{if } \theta \le 1 - 1/(\lambda-1), \\ (1 + \theta f(1-\theta)) \, |\mathbf{E}|^2 & \text{if } \theta \ge 1 - 1/(\lambda-1). \end{cases} \tag{1.8}$$

The upper bound is attained by a suitable extremal sequence of configurations in ad_θ. The lower bound is attained by a configuration made up of parallel layers of the β conductor with layer normals orthogonal to $\mathbf{E}$. These results hold for all bounded domains Ω of R^N, $N \ge 2$ with Lipschitz boundary.

Extremal sequences of configurations that attain the upper bound are found to be given by the well-known finite rank laminar microstructures. This class of configurations is known to give extremal effective conductivity properties; see [7] and [9]. They also arise in the study of minimization problems for integral functionals of the form $\int_\Omega W(\nabla \phi) \, dx$ with nonconvex energy densities W; see [1], [3], [4], and [6].

It is shown here that only laminates of the first and second rank appear in extremal sequences of configurations. In order to describe a second rank laminate we introduce two characteristic functions, one for each scale of oscillation. We consider the periodic function $\chi^1(t)$ defined on the real line with period $0 \le t \le 1$ such that $\chi^1 = 1$ for $0 \le t \le \theta_1$ and $\chi^1 = 0$ elsewhere. Similarly we introduce the unit periodic function χ^2 such that $\chi^2 = 1$ for $0 \le t \le \theta_2$ and $\chi^2 = 0$ elsewhere. We introduce unit vectors $\mathbf{n}^1$ and $\mathbf{n}^2$ representing layer directions and put

$$\chi_L^\varepsilon(x) = \left(1 - \chi^1\left(\frac{\mathbf{n}^1 \cdot x}{\varepsilon}\right)\right)\left(1 - \chi^2\left(\frac{\mathbf{n}^2 \cdot x}{\varepsilon^2}\right)\right). \tag{1.9}$$

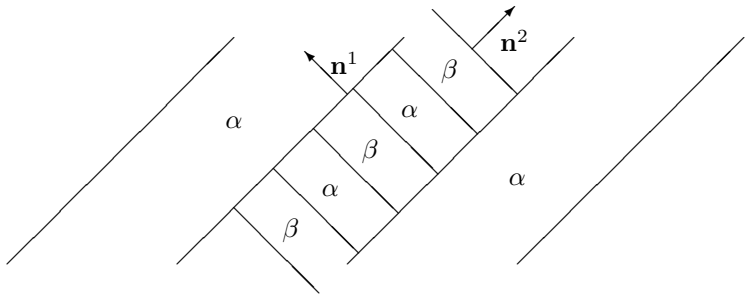


FIG. 1. *A laminate of second rank.*

The configurations associated with the sequence of characteristic functions $\{\chi_L^\varepsilon\}_{\varepsilon>0}$ are referred to as a laminate of the second rank. The conductivities for this sequence of configurations are given by $a(\chi_L^\varepsilon)$; see Figure 1. The laminate of first rank has one less scale of oscillation and is given by

$$\chi_L^\varepsilon(x) = \left(1 - \chi^1\left(\frac{\mathbf{n}^1 \cdot x}{\varepsilon}\right)\right). \tag{1.10}$$

The sequence of temperature gradients associated with laminates of rank one or two is written as $\{\nabla T_L^\varepsilon\}_{\varepsilon>0}$, where

$$-\text{div } (a(\chi_L^\varepsilon)\nabla T_L^\varepsilon) = 0 \tag{1.11}$$

and $T_L^\varepsilon = \mathbf{E}\cdot\mathbf{x}$ on the boundary of Ω.

In general, we may consider any sequence of configurations $\{\chi^\varepsilon\}_{\varepsilon>0}$ indexed by ε and the associated sequence of temperature gradients $\{\nabla T^\varepsilon\}_{\varepsilon>0}$ satisfying $-\text{div } (a(\chi^\varepsilon)\nabla T^\varepsilon) = 0$ and $T^\varepsilon = \mathbf{E}\cdot\mathbf{x}$ on the boundary of Ω. A sequence of configurations $\{\chi^\varepsilon\}_{\varepsilon>0}$ is said to be a maximizing sequence if

$$\lim_{\varepsilon\to 0} \|\nabla T^\varepsilon\|^2 = U(\theta, \mathbf{E}). \tag{1.12}$$

A configuration is said to be minimizing if the associated temperature T satisfies

$$\|\nabla T\|^2 = meas(\Omega) \times |\mathbf{E}|^2. \tag{1.13}$$

The next result identifies configurations that minimize the L^2 norm of the gradient.

THEOREM 1.2 (minimizing configurations for the L^2 norm). *Given* $\mathbf{E}$ *in* R^N, *a minimizing configuration is obtained by placing the* β *conductor in layers oriented so that the layer normals are orthogonal to* $\mathbf{E}$. *The number and thickness of the layers is constrained only by the requirement that the configuration be in* ad_θ.

It is easily shown that the temperature for this configuration is given by $T = \mathbf{E}\cdot\mathbf{x}$ everywhere in Ω; see section 4.

We now identify maximizing sequences of configurations.

THEOREM 1.3 (maximizing sequences of configurations for the L^2 norm). *Given* $\mathbf{E}$ *in* R^N

1. *If $\lambda \leq 2$, then a maximizing sequence of configurations is given by a laminate of the first rank in ad_θ with layer normal $\mathbf{n}^1$ parallel to $\mathbf{E}$ and $\theta^1 = 1 - \theta$.*
2. *If $\lambda > 2$, then*
   (a) *if $\theta \leq 1 - 1/(\lambda - 1)$, then a maximizing sequence of configurations is given by a laminate of the second rank in ad_θ with layer normal $\mathbf{n}^1$ parallel to $\mathbf{E}$, layer normal $\mathbf{n}^2$ orthogonal to $\mathbf{E}$, $\theta_1 = \frac{1}{1+\theta(\lambda-1)}$, and $\theta_2 = 1 - \theta - (\frac{1}{\lambda-1})$;*
   (b) *if $\theta \geq 1 - 1/(\lambda - 1)$, then a maximizing sequence of configurations is given by a laminate of the first rank in ad_θ with layer normal $\mathbf{n}^1$ parallel to $\mathbf{E}$ and $\theta_1 = 1 - \theta$.*

The geometry for minimizing configurations is independent of the volume fraction of the β phase and the contrast λ. On the other hand, for $\lambda > 2$, we see that the maximizing sequences of configurations change from laminates of rank one to laminates of rank two when the volume fraction of the β phase drops below $1 - \frac{1}{\lambda-1}$. When this happens the extremal configuration of α and β phases changes topology and the α phase occupies a connected set, while the β phase is in the form of thin rectangular inclusions.

In view of the applications, it is important to control the temperature gradient, as regions containing large temperature gradients are most often the first to suffer damage during service. Theorem 1.2 provides rigorous rules of thumb for the design of configurations for minimizing the L^2 norm of the temperature gradient, i.e., minimizing configurations are given by layering the two conductors in strips parallel to the applied field $\mathbf{E}$. On the other hand, the upper bound given in Theorem 1.1 provides the best possible upper bound on the L^2 norm of the temperature gradient when only the volume fraction of the β phase is known. We point out that the upper bound goes to infinity with the contrast λ.

The basic idea behind our approach is to encode the constraint given by the equilibrium condition (1.1) directly into the cost functional $\|\nabla T\|_2^2$. To do this we follow Golden and Papanicolaou [5] and introduce a scattering theory formalism to express ∇T in terms of the solution operator for (1.1). We then substitute the representation for ∇T into the L^2 norm to obtain the desired Stieltjes representation for $\|\nabla T\|_2^2$ in terms of a matrix valued measure. Using perturbation theory we see as in [5] that there are an infinite number of constraints on the matrix valued measure. We judiciously choose a subset of these constraints associated with the first and second moments of the measure. Our choice is motivated by the corrector theory of homogenization for laminates of finite rank given by Briane [2]. Subject to these constraints we extremize the Stietljes representation formula over all associated matrix valued measures to obtain the bounds given in Theorem 1.1. The attainability of the upper bound is established by comparing it to the limits of the L^2 norms associated with laminates of rank one or two. The comparison is facilitated using an explicit Stieltjes integral representation formula for these limits. We are confident that the approach developed here will be successful for investigating analogous problems in the elasticity setting.

The paper is organized as follows. In section 2 we review the recent results of Briane [2] that give the explicit form of corrector matrices for laminates of finite rank. We apply this theory to write the limit of the L^2 norms for finite rank laminates as Stieltjes functions. In section 3 we develop a Stieltjes representation formula for the L^2 norm of the gradient for any admissible configuration. In section 4 we use the representation formula to obtain the bounds stated in Theorem 1.1 and to establish their optimality.

2. Correctors and the L^2 norm for layered materials. In this section we obtain an explicit formula for

$$\lim_{\varepsilon\to 0}\|\nabla T_L^\varepsilon\|_2^2.$$

We start by reviewing the notion of H convergence due to Spagnolo [10] and Murat and Tartar [8]. We consider the sequence of conductivities $\{a(\chi^\varepsilon)\}_{\varepsilon>0}$ associated with the sequence of configurations $\{\chi^\varepsilon\}_{\varepsilon>0}$. The sequence $\{a(\chi^\varepsilon)\}_{\varepsilon>0}$ is said to H converge to A if for any function f of $H^{-1}(\Omega)$ the solutions $u^\varepsilon \in W_0^{1,2}(\Omega)$ of

$$-\text{div }(a(\chi^\varepsilon)\nabla u^\varepsilon) = f$$

satisfy $u^\varepsilon \to u^0$ weakly in $W^{1,2}(\Omega)$ and $a(\chi^\varepsilon)\nabla u^\varepsilon \to A\nabla u^0$ weakly in $L^2(\Omega; R^N)$, where u^0 is the solution of $-\text{div }\left(A\nabla u^0\right) = f$ and $u^0 \in W_0^{1,2}(\Omega)$. In fact more can be said about the convergence of the sequence $\{\nabla u^\varepsilon\}_{\varepsilon>0}$. There exists a matrix field $\mathbf{P}^\varepsilon$, called a corrector, for which

$$\nabla u^\varepsilon = \mathbf{P}^\varepsilon \nabla u^0 + z^\varepsilon,$$

where $z^\varepsilon \to 0$ strongly in $L^1(\Omega; R^N)$. Tartar [11] and Murat and Tartar [8] prove there always exists such a sequence of correctors $\mathbf{P}^\varepsilon$.

We choose layering directions $\mathbf{n}^1$ and $\mathbf{n}^2$ so that they are orthogonal to each other and put $\chi_1^\varepsilon = \chi^1(\frac{\mathbf{n}^1\cdot\mathbf{x}}{\varepsilon})$ and $\chi_2^\varepsilon = \chi^2(\frac{\mathbf{n}^2\cdot\mathbf{x}}{\varepsilon^2})$. We invoke Theorem 2.1 of Briane [2], and a straightforward calculation shows that the correctors are of the form

$$\mathbf{P}^\varepsilon = \chi_1^\varepsilon \mathbf{P}^1 + (1-\chi_1^\varepsilon)\left[\chi_2^\varepsilon \mathbf{P}^2 + (1-\chi_2^\varepsilon)\mathbf{P}^3\right], \tag{2.1}$$

where the constant matrices $\mathbf{P}^1$, $\mathbf{P}^2$, and $\mathbf{P}^3$ are given by

$$\mathbf{P}^1 = \mathbf{I} + (1-\theta_1)\left(\frac{(1-\theta_2)(\lambda-1)}{1-\theta_1(1-\theta_2)+\theta_1(1-\theta_2)\lambda}\right)\mathbf{n}^1\otimes\mathbf{n}^1, \tag{2.2}$$

$$\mathbf{P}^2 = \mathbf{I} - \theta_1\left(\frac{(1-\theta_2)(\lambda-1)}{1-\theta_1(1-\theta_2)+\theta_1(1-\theta_2)\lambda}\right)\mathbf{n}^1\otimes\mathbf{n}^1 + (1-\theta_2)\left(\frac{\lambda-1}{(1-\theta_2)+\theta_2\lambda}\right)\mathbf{n}^2\otimes\mathbf{n}^2, \tag{2.3}$$

and

$$\mathbf{P}^3 = \mathbf{I} - \theta_1\left(\frac{(1-\theta_2)(\lambda-1)}{1-\theta_1(1-\theta_2)+\theta_1(1-\theta_2)\lambda}\right)\mathbf{n}^1\otimes\mathbf{n}^1 - \theta_2\left(\frac{\lambda-1}{(1-\theta_2)+\theta_2\lambda}\right)\mathbf{n}^2\otimes\mathbf{n}^2, \tag{2.4}$$

where $\mathbf{I}$ is the $N\times N$ identity and $\mathbf{n}^1\otimes\mathbf{n}^1$ and $\mathbf{n}^2\otimes\mathbf{n}^2$ are the rank one matrices $\mathbf{n}_i^1\mathbf{n}_j^1$ and $\mathbf{n}_i^2\mathbf{n}_j^2$, respectively. The H limit for the sequence $\{a(\chi_L^\varepsilon)\}_{\varepsilon>0}$ is a constant $N\times N$ matrix denoted by A^L [7], [8]. For the boundary value problem treated here we have that

$$\begin{aligned} T^\varepsilon &\to \mathbf{E}\cdot\mathbf{x} \quad \text{weakly in } W^{1,2}(\Omega) \text{ and} \\ a(\chi^\varepsilon)\nabla T^\varepsilon &\to A^L\mathbf{E} \quad \text{weakly in } L^2(\Omega; R^N). \end{aligned} \tag{2.5}$$

From the corrector theory we have

$$\nabla T_L^\varepsilon = \mathbf{P}^\varepsilon\mathbf{E} + z^\varepsilon. \tag{2.6}$$

It is evident from the formulas describing $\mathbf{P}^\varepsilon$ that the sequence $\{\mathbf{P}^\varepsilon\}_{\varepsilon>0}$ is uniformly bounded in $L^\infty(\Omega;\, R^{N\times N})$. Thus we appeal to Theorem 3 of Murat and Tartar [8] to find that $z^\varepsilon \to 0$ strongly in $L^2(\Omega;\, R^N)$. We note that because of the separation of scales, the sequence of products $\{\chi_1^\varepsilon \chi_2^\varepsilon\}_{\varepsilon>0}$ converges in a weak L^∞ *star* to the product $\theta_1\theta_2$. Collecting our results and taking limits we find that

$$\lim_{\varepsilon\to 0} \|\nabla T_L^\varepsilon\|_2^2 = \mathbf{C}_{ij}^L \mathbf{E}_i \mathbf{E}_j, \tag{2.7}$$

where the matrix $\mathbf{C}^L$ is given by

$$\begin{aligned}\mathbf{C}^L &= meas(\Omega)\mathbf{I} \\ &\quad + meas(\Omega)(1-\theta_1)(1-\theta_2)\left(\frac{\theta_1(1-\theta_2)}{(\frac{1}{1-\lambda}-\theta_1(1-\theta_2))^2}\right)\mathbf{n}^1\otimes\mathbf{n}^1 \\ &\quad + meas(\Omega)(1-\theta_1)(1-\theta_2)\left(\frac{\theta_2}{(\frac{1}{1-\lambda}-\theta_2)^2}\right)\mathbf{n}^2\otimes\mathbf{n}^2. \end{aligned}\tag{2.8}$$

Here we note that the total volume fraction of the β phase is given by $\theta = (1-\theta_1)(1-\theta_2)$, and we can rewrite (2.8) as

$$\mathbf{C}^L = \mathbf{C}(\boldsymbol{\mu}^L) = meas(\Omega)\times\left(\mathbf{I} + \int_0^{1-\theta} f(z)\boldsymbol{\mu}^L(dz)\right), \tag{2.9}$$

where the matrix valued measure $\boldsymbol{\mu}$ is given by

$$\boldsymbol{\mu}^L(dz) = \left(\theta\delta(z-\theta_1(1-\theta_2))\mathbf{n}^1\otimes\mathbf{n}^1 + \theta\delta(z-\theta_2)\mathbf{n}^2\otimes\mathbf{n}^2\right)dz. \tag{2.10}$$

Equations (2.9) and (2.10) provide the desired Stieltjes integral formula for the limit given in (2.7).

3. Stieltjes integral representation formula for the L^2 norm. In this section we develop a representation formula for $\|\nabla T\|_2^2$, where T is the solution of (1.1) and $T = \mathbf{E}\cdot\mathbf{x}$ on the boundary of Ω. Motivated by perturbation theory we shall first rewrite the constraint given by (1.1). To this end we introduce the solution operator $(-\Delta)^{-1}$ mapping $H^{-1}(\Omega)$ onto $W_0^{1,2}(\Omega)$ for the problem given by $w \in W_0^{1,2}(\Omega)$ and

$$-\Delta w = f \quad \text{on } \Omega. \tag{3.1}$$

Next we introduce the subspace $\mathcal{E}$ of $L^2(\Omega;\, R^N)$ defined by

$$\mathcal{E} = \{\eta \in L^2(\Omega;\, R^N) \mid \eta = \nabla\varphi,\ \varphi \in W_0^{1,2}(\Omega)\},$$

and we introduce the operator P defined by $P = \partial_{x_i}(\Delta)^{-1}\partial_{x_j}$. It is easily checked that the operator P is a projection from $L^2(\Omega;\, R^N)$ into the subspace $\mathcal{E}$. We introduce the field perturbation $\phi = T - \mathbf{E}\cdot\mathbf{x}$ and rewrite the conductivity $a(\chi)$ as a positive perturbation from the uniform state α, i.e., $a(\chi) = \alpha + (\beta-\alpha)\chi$. Expanding T and $a(\chi)$ in (1.1) gives

$$-\alpha\Delta\phi = \text{div}\,\left((\beta-\alpha)\chi(\nabla\phi + \mathbf{E})\right). \tag{3.2}$$

Dividing both sides by α, applying $(-\Delta)^{-1}$ to both sides, and manipulating gives

$$\nabla\phi + \mathbf{E} + P\left[\,(\lambda-1)\chi\,(\nabla\phi+\mathbf{E})\,\right] = \mathbf{E} \tag{3.3}$$

or

$$[\,\mathbf{I} + (\lambda - 1)\,\Lambda\,]\,\nabla T = \mathbf{E}, \tag{3.4}$$

where $\Lambda = P\chi$. From this we obtain the desired expression

$$\nabla T = [\,\mathbf{I} + (\lambda - 1)\,\Lambda\,]^{-1}\,\mathbf{E}. \tag{3.5}$$

It is clear that the equilibrium constraint (1.1) is now explicitly encoded in the formula for ∇T as given by (3.5). The next step is to rewrite $\|\nabla T\|_2^2$ in a way that exploits the spectral properties of the operator Λ. To do this we expand the energy dissipation denoted by Q in two different ways. Here

$$Q = \frac{1}{meas(\Omega)} \int_\Omega a(\chi)\,\nabla T \cdot \nabla T\,dx. \tag{3.6}$$

Expanding $a(\chi)$ as $a(\chi) = \alpha + \chi(\beta - \alpha)$ and substitution into (3.6) gives

$$Q = \frac{1}{meas(\Omega)} \int_\Omega \alpha |\nabla T|^2\,dx + \frac{(\beta - \alpha)}{meas(\Omega)} \int_\Omega \chi\,|\nabla T|^2\,dx. \tag{3.7}$$

We expand ∇T as $\nabla T = \nabla\phi + \mathbf{E}$ in (3.6) to obtain

$$\begin{aligned} Q &= \frac{1}{meas(\Omega)} \int_\Omega a(\chi)\,\nabla T \cdot \mathbf{E}\,dx \\ &= \alpha |\mathbf{E}|^2 + \frac{(\beta - \alpha)}{meas(\Omega)} \int_\Omega \chi\,\nabla T \cdot \mathbf{E}\,dx. \end{aligned} \tag{3.8}$$

Here the first equality in (3.8) follows from (1.2), and the second follows from expansion of $a(\chi)$ and $\int_\Omega \nabla\phi \cdot \mathbf{E}\,dx = 0$. Eliminating Q using (3.7) and (3.8) gives

$$\|\nabla T\|_2^2 = meas(\Omega)\left(|\mathbf{E}|^2 + \frac{(\lambda - 1)}{meas(\Omega)} \int_\Omega \chi\,(\nabla T \cdot \mathbf{E})\,dx - \frac{(\lambda - 1)}{meas(\Omega)} \int_\Omega \chi\,|\nabla T|^2\,dx\right). \tag{3.9}$$

For vector fields η and ψ in $L^2(\Omega;\, R^N)$ we introduce the bilinear form $\langle \eta, \psi \rangle$ defined by

$$\langle \eta, \psi \rangle = \frac{1}{meas(\Omega)} \int_\Omega \chi\,(\eta \cdot \psi)\,dx,$$

and $\langle \eta, \psi \rangle$ is an inner product for the Hilbert space $\mathcal{H}$ defined by

$$\mathcal{H} = \left\{ \psi \in L^2(\Omega;\, R^N) \text{ modulo the equivalence class of elements } \psi \right.$$
$$\left. \text{such that } \int_\Omega \chi\psi\,dx = 0 \right\}.$$

Substitution of (3.5) into (3.9) gives

$$\begin{aligned} \|\nabla T\|_2^2 &= meas(\Omega)|\mathbf{E}|^2 \\ &\quad + meas(\Omega)(\lambda - 1)\left\langle [\,\mathbf{I} + (\lambda - 1)\,\Lambda\,]^{-1}\,\mathbf{E}, \mathbf{E} \right\rangle \\ &\quad - meas(\Omega)(\lambda - 1)\left\langle [\,\mathbf{I} + (\lambda - 1)\,\Lambda\,]^{-1}\,\mathbf{E}, [\,\mathbf{I} + (\lambda - 1)\,\Lambda\,]^{-1}\,\mathbf{E} \right\rangle. \end{aligned} \tag{3.10}$$

It is easily seen that Λ is a positive symmetric operator on $\mathcal{H}$ with norm less than or equal to 1. From spectral theory we immediately obtain the existence of a projection valued measure $R(dz)$ with support on $[0,1]$ for which

$$\left\langle [\,\mathbf{I} + (\lambda-1)\,\Lambda\,]^{-1}\,\mathbf{E},\mathbf{E}\right\rangle = \left\langle \int_0^1 \frac{1}{1+z(\lambda-1)}\,R(dz)\mathbf{E},\mathbf{E}\right\rangle \tag{3.11}$$

and

$$\left\langle [\,\mathbf{I} + (\lambda-1)\,\Lambda\,]^{-1}\,\mathbf{E},[\,\mathbf{I} + (\lambda-1)\,\Lambda\,]^{-1}\,\mathbf{E}\right\rangle = \left\langle \int_0^1 \frac{1}{(1+z(\lambda-1))^2}\,R(dz)\mathbf{E},\mathbf{E}\right\rangle. \tag{3.12}$$

Collecting our results we arrive at the Stieltjes integral representation formula given by the following theorem.

THEOREM 3.1 (Stieltjes integral representation formula).

$$\|\nabla T\|_2^2 = \mathbf{C}_{ij}(\boldsymbol{\mu})\mathbf{E}_i\mathbf{E}_j, \tag{3.13}$$

where

$$\mathbf{C}(\boldsymbol{\mu}) = meas(\Omega)\left(\mathbf{I} + \int_0^1 f(z)\,\boldsymbol{\mu}(dz)\right) \tag{3.14}$$

and

$$\boldsymbol{\mu}_{ij}(dz) = \langle R(dz)\,\mathbf{e}^i,\mathbf{e}^j\rangle. \tag{3.15}$$

Here $\mathbf{e}^i$, $i=1,2\ldots,N$ *is an orthonormal basis for* R^N. *Moreover,* $\boldsymbol{\mu}_{ij}=\boldsymbol{\mu}_{ji}$, *since* $R(dz)$ *is symmetric and for all* $\mathbf{E}$ *in* R^N *we have that the measures* $\boldsymbol{\mu}(dz)\mathbf{E}\cdot\mathbf{E}$ *are positive.*

It is evident from Theorem 3.1 that the geometric information is stored in the measure $\boldsymbol{\mu}$ while the ratio of conductivities is contained in $f(z)$. The extremal behavior of $\|\nabla T\|_2^2$ is governed by the global maxima and minima of f on $[0,1]$.

4. Derivation of the isoperimetric inequalities. In view of the Stieltjes formula for the gradient we can replace the extremal problems

$$\mathrm{A} = \inf_{a(\chi)\in ad_\theta}\left\{\|\nabla T\|_2^2\,;\,-\mathrm{div}\,(a(\chi)\nabla T)=0,\ T=\mathbf{E}\cdot\mathbf{x}\ \text{on}\ \partial\Omega\right\} \tag{4.1}$$

and

$$\mathrm{B} = \sup_{a(\chi)\in ad_\theta}\left\{\|\nabla T\|_2^2\,;\,-\mathrm{div}\,(a(\chi)\nabla T)=0,\ T=\mathbf{E}\cdot\mathbf{x}\ \text{on}\ \partial\Omega\right\} \tag{4.2}$$

with the equivalent problems given by

$$\mathrm{A} = \inf_{\boldsymbol{\mu}\in\mathcal{A}_\theta}\{\mathbf{C}(\boldsymbol{\mu})\mathbf{E}\cdot\mathbf{E}\} \tag{4.3}$$

and

$$\mathrm{B} = \sup_{\boldsymbol{\mu}\in\mathcal{A}_\theta}\{\mathbf{C}(\boldsymbol{\mu})\mathbf{E}\cdot\mathbf{E}\}. \tag{4.4}$$

Here the set $\mathcal{A}_\theta$ is the set of measures $\boldsymbol{\mu}$ given by (3.15) associated with any configuration of the β phase described by a characteristic function χ subject to the isoperimetric

constraint $\int_\Omega \chi\, dx = \theta\, meas(\Omega)$. Instead of attempting an explicit characterization of $\mathcal{A}_\theta$ we introduce a larger set of measures $\overline{\mathcal{A}_\theta}$ and compute the lower and upper bounds

$$\underline{\mathrm{A}} = \inf_{\boldsymbol{\mu}\in\overline{\mathcal{A}_\theta}} \{\mathbf{C}(\boldsymbol{\mu})\mathbf{E}\cdot\mathbf{E}\} \tag{4.5}$$

and

$$\overline{\mathrm{B}} = \sup_{\boldsymbol{\mu}\in\overline{\mathcal{A}_\theta}} \{\mathbf{C}(\boldsymbol{\mu})\mathbf{E}\cdot\mathbf{E}\}\,. \tag{4.6}$$

The goal here is to find a suitable choice for $\overline{\mathcal{A}_\theta}$ that delivers optimal bounds. We start by examining constraints on the measure $\boldsymbol{\mu}^L(dz)$ associated with laminates of the second rank defined in (2.10). One readily sees that

$$\int_0^1 \boldsymbol{\mu}^L(dz) = \theta\mathbf{I}, \tag{4.7}$$

and since $1-\theta = \theta_2 + \theta_1(1-\theta_2)$ we have

$$\mathbf{T}^L \triangleq \int_0^1 z\,\boldsymbol{\mu}^L(dz) \le (\max\{\theta\theta_1(1-\theta_2), \theta\theta_2\}) \times \mathbf{I} \le \theta(1-\theta)\mathbf{I}. \tag{4.8}$$

Next, for comparison, we apply perturbation expansions to look for constraints on $\boldsymbol{\mu}(dz)$. Expansion about $\lambda = 1$ gives

$$[\,\mathbf{I} + (\lambda-1)\,\Lambda\,]^{-1} = \mathbf{I} + (1-\lambda)\,\Lambda + (1-\lambda)^2\,\Lambda^2 + (1-\lambda)^3\,\Lambda^3 + \cdots \tag{4.9}$$

and

$$\frac{1}{1+z(\lambda-1)} = 1 + (1-\lambda)\,z + (1-\lambda)^2\,z^2 + (1-\lambda)^3\,z^3 + \cdots. \tag{4.10}$$

Substituting these expansions into (3.11) and equating like powers of $\lambda - 1$ gives

$$\int_0^1 z^n\,\boldsymbol{\mu}_{ij}(dz) = \langle \Lambda^n \mathbf{e}^i, \mathbf{e}^j\rangle, \quad n = 0, 1\ldots. \tag{4.11}$$

Focusing on the cases $n = 0$ and $n = 1$ we have

$$\int_0^1 \boldsymbol{\mu}_{ij}(dz) = \theta\,\mathbf{I}_{ij} \tag{4.12}$$

and

$$\int_0^1 z\,\boldsymbol{\mu}_{ij}(dz) = \langle \Lambda\mathbf{e}^i, \mathbf{e}^j\rangle. \tag{4.13}$$

We estimate the largest and smallest eigenvalues for the tensor $\mathbf{T}_{ij} \triangleq \langle \Lambda\mathbf{e}^i, \mathbf{e}^j\rangle$. We note that constant vectors lie in the null space of the operator P, and we introduce $\overline{\chi} = \chi - \theta$ to find that

$$\begin{aligned} 0 \le \mathbf{T}_{ij}\mathbf{E}_i\mathbf{E}_j &= \frac{1}{meas(\Omega)}\int_\Omega (P\chi)\mathbf{E}\cdot\chi\mathbf{E}\,dx \\ &= \frac{1}{meas(\Omega)}\int_\Omega (P\overline{\chi})\mathbf{E}\cdot\overline{\chi}\mathbf{E}\,dx \\ &\le \frac{1}{meas(\Omega)}\int_\Omega (\overline{\chi})^2\,dx|\mathbf{E}|^2 = \theta(1-\theta)|\mathbf{E}|^2. \end{aligned} \tag{4.14}$$

Thus the spectrum of the tensor $\mathbf{T}$ lies in the interval $[0, \theta(1-\theta)]$. Motivated by (4.7), (4.8), (4.12), and (4.14) we define $\overline{\mathcal{A}_\theta}$ to be given by all $N \times N$ symmetric matrices with elements given by finite Borel measures such that for any vector $\mathbf{v}$ the measure given by $\boldsymbol{\mu}(dz)\mathbf{v}\cdot\mathbf{v}$ is positive and the matrix of measures satisfies the moment constraints

$$\int_0^1 \boldsymbol{\mu}(dz) = \theta\,\mathbf{I} \tag{4.15}$$

and

$$\int_0^1 z\,\boldsymbol{\mu}(dz) = \mathbf{T}, \tag{4.16}$$

where $\mathbf{T}$ is a symmetric $N \times N$ matrix with eigenvalues contained in the interval $[0, \theta(1-\theta)]$. Its clear from the definition of $\overline{\mathcal{A}_\theta}$ that this set of measures contains $\mathcal{A}_\theta$.

For the purpose of computing the bounds $\underline{\mathrm{A}}$ and $\overline{\mathrm{B}}$ we characterize the range of the map $\mathbf{H}(\boldsymbol{\mu})$ given by

$$\mathbf{H}(\boldsymbol{\mu}) = \int_0^1 f(z)\,\boldsymbol{\mu}(dz) \tag{4.17}$$

for $\boldsymbol{\mu}$ in $\overline{\mathcal{A}_\theta}$. We introduce the set $\overline{\mathcal{V}}_\theta$ of vectors $(\nu_1, \nu_2, \ldots, \nu_N)$ whose elements are positive finite Borel measures supported on $[0,1]$ that satisfy the constraints

$$\int_0^1 \nu_i(dz) = \theta, \quad i = 1, \ldots, N, \tag{4.18}$$

and

$$\int_0^1 z\,\nu_i(dz) = m_i, \quad \text{where } 0 \le m_i \le \theta(1-\theta), \quad i = 1, \ldots, N. \tag{4.19}$$

We now state the following theorem.

THEOREM 4.1 (the matrix range of $\mathbf{H}(\boldsymbol{\mu})$). *Let* R *be the image of* $\overline{\mathcal{A}_\theta}$ *under the map* $\mathbf{H} : \overline{\mathcal{A}_\theta} \to R^{N\times N}$. *Then* R *is given by*

$$\mathrm{R} = \left\{ \begin{array}{c} M \in R^{N\times N}\,;\ M = \Sigma_{i=1}^N \lambda_i \mathbf{e}^i \otimes \mathbf{e}^i, \\ \textit{where } \lambda_i = \int_0^1 f(z)\,\nu_i(dz), \textit{and } (\nu_1, \nu_2\ \ldots \nu_N) \textit{ in } \overline{\mathcal{V}}_\theta, \\ \textit{and } \mathbf{e}^i,\ i = 1, \ldots, N, \quad \textit{is any orthonormal basis for } R^N. \end{array} \right\}. \tag{4.20}$$

Proof. We denote the right-hand side of (4.20) by S and show $R = S$. One sees that $S \subset R$ by writing $M = \int_0^1 f(z)\,P(dz)$, where $P(dz) = \Sigma_i \nu_i(dz)\mathbf{e}^i \otimes \mathbf{e}^i$, and checking (4.15) and (4.16). To show $R \subset S$ we consider the matrix M given by $M = \int_0^1 f(z)\boldsymbol{\mu}(dz)$. Since M is symmetric it has an orthonormal system of eigenvectors $\mathbf{v}^i$, $i = 1 \ldots, N$, and $M = \Sigma_i \lambda_i \mathbf{v}^i \otimes \mathbf{v}^i$. From this one deduces that $\lambda_i = \int_0^1 f(z)\mu_i(dz)$, where μ_i are the positive measures given by $\mu_i(dz) = \boldsymbol{\mu}(dz)\mathbf{v}^i \cdot \mathbf{v}^i$. Next we observe that

$$\int_0^1 \mu_i(dz) = \int_0^1 \boldsymbol{\mu}(dz)\mathbf{v}^i \cdot \mathbf{v}^i = \theta \tag{4.21}$$

and

$$m_i = \int_0^1 z\,\mu_i(dz) = \int_0^1 z\,\boldsymbol{\mu}(dz)\mathbf{v}^i \cdot \mathbf{v}^i \le \theta(1-\theta) \tag{4.22}$$

to discover that $(\mu_1, \dots \mu_N)$ lies in $\overline{\mathcal{V}}_\theta$, and the theorem is proved. ∎

We now establish the explicit formulas for the bounds given by the following theorem.

THEOREM 4.2 (bounds on the L^2 norm of the gradient).

$$\underline{\mathrm{A}} = meas(\Omega)|\mathbf{E}|^2 \le \|\nabla T\|_2^2 \le \overline{\mathrm{B}} = U(\theta, \mathbf{E}). \tag{4.23}$$

Before establishing the theorem we note that the lower bound $meas(\Omega)|\mathbf{E}|^2$ can be found directly. Indeed, we can write $T = \phi + \mathbf{E}\cdot\mathbf{x}$, where $\phi = 0$ on the boundary of Ω. Then expanding $\|\nabla T\|_2^2$ and noting that $\int_\Omega \nabla\phi\cdot\mathbf{E}\,dx = 0$, we have

$$\|\nabla T\|_2^2 = meas(\Omega)|\mathbf{E}|^2 + \int_\Omega |\nabla\phi|^2\,dx, \tag{4.24}$$

and the lower bound follows immediately.

Proof of Theorem 4.2. We start by proving $\underline{\mathrm{A}} = meas(\Omega)|\mathbf{E}|^2$. From Theorem 4.1 it follows that

$$\underline{\mathrm{A}} = \inf_{M\in R} \left\{ meas(\Omega)(\mathbf{I}+M)\,\mathbf{E}\cdot\mathbf{E}\right\}. \tag{4.25}$$

It is evident from (4.25) that for $(\nu_1, \dots, \nu_N)$ fixed the minimum occurs when $\mathbf{E}$ lies in the eigenspace of the smallest eigenvalue of M. Without loss of generality we assume that $\lambda_1 = \int_0^1 f(z)\,\nu_1(dz)$ is the smallest eigenvalue of M, and we choose $\mathbf{e}^1 = \mathbf{E}/|\mathbf{E}|$ to find that

$$\underline{\mathrm{A}} = \inf_{\substack{\nu_1\ge 0,\\ \int_0^1 \nu_1(dz)=\theta, \int_0^1 z\nu_1(dz)\le\theta(1-\theta)}} \left\{ meas(\Omega)\left(1+\int_0^1 f(z)\,\nu_1(dz)\right)|\mathbf{E}|^2\right\}. \tag{4.26}$$

To finish the minimization we note that for $\lambda > 1$ the function $f(z)$ is strictly positive on $0 < z < \infty$ with $f(0) = 0$ and $\lim_{z\to\infty} f(z) = 0$. Moreover, $f(z)$ has a global maximum over $[0,\infty)$ at $z = 1/(\lambda-1)$ with $\acute{f}(z) \ge 0$ for $z \le 1/(\lambda-1)$ and $\acute{f}(z) \le 0$ for $z \ge 1/(\lambda-1)$. With this in mind we choose $\nu_1(dz) = \theta\delta(z)dz$. Since this choice is admissible we have established that $\underline{\mathrm{A}} = meas(\Omega)|\mathbf{E}|^2$.

Next we establish the upper bound. From Theorem 4.1 it follows that

$$\overline{\mathrm{B}} = \sup_{M\in R} \left\{ meas(\Omega)(\mathbf{I}+M)\,\mathbf{E}\cdot\mathbf{E}\right\}. \tag{4.27}$$

Here it is evident that for $(\nu_1, \dots, \nu_N)$ fixed the maximum occurs when $\mathbf{E}$ lies in the eigenspace of the largest eigenvalue of M. Without loss of generality we assume that $\lambda_1 = \int_0^1 f(z)\,\nu_1(dz)$ is the largest eigenvalue of M, and we choose $\mathbf{e}^1 = \mathbf{E}/|\mathbf{E}|$ to find that

$$\overline{\mathrm{B}} = \sup_{\substack{\nu_1\ge 0,\\ \int_0^1 \nu_1(dz)=\theta, \int_0^1 z\nu_1(dz)\le\theta(1-\theta)}} \left\{ meas(\Omega)\left(1+\int_0^1 f(z)\,\nu_1(dz)\right)|\mathbf{E}|^2\right\}. \tag{4.28}$$

To proceed we normalize and write $\nu_1(dz) = \theta p(dz)$. The extremal problem becomes

$$\overline{\mathrm{B}} = meas(\Omega)\left(|\mathbf{E}|^2 + \theta\sup_{p\in\mathcal{C}}\left(\int_0^1 f(z)\,p(dz)\right)|\mathbf{E}|^2\right), \tag{4.29}$$

where $\mathcal{C}$ is the set of probability measures for which $0 \leq \bar{z} = \int_0^1 z\, p(dz) \leq (1-\theta)$. Noting that the function $f(z)$ is strictly concave over an interval that includes $[0, 1/(\lambda-1)]$, strictly increasing on $[0, 1/(\lambda-1)]$, and strictly decreasing on $(1/(\lambda-1), \infty)$, we have for $(1-\theta) \leq 1/(\lambda-1)$ that

$$\int_0^1 f(z)\, p(dz) < f(\bar{z}) \leq f(1-\theta). \tag{4.30}$$

It is evident that the best choice is $p(dz) = \delta(z-(1-\theta))$, and we find that

$$\overline{\mathrm{B}} = meas(\Omega) \times (1 + \theta f(1-\theta))\, |\mathbf{E}|^2. \tag{4.31}$$

We observe that for $1 < \lambda \leq 2$ we have $1/(\lambda-1) \geq 1$ and for $0 \leq \theta \leq 1$ we always have $(1-\theta) \leq 1/(\lambda-1)$. On the other hand, when $(1-\theta) \geq 1/(\lambda-1)$ it is evident that the best choice corresponds to the global maximum of f, i.e., $p(dz) = \delta(z - 1/(\lambda-1))$, and we find that

$$\overline{\mathrm{B}} = meas(\Omega) \times (1 + \theta f(1/(\lambda-1)))\, |\mathbf{E}|^2, \tag{4.32}$$

and the theorem follows. □

We conclude by proving Theorems 1.2 and 1.3. To prove Theorem 1.2 we show that the lower bound $\underline{\mathrm{A}}$ is attained by configurations made up of layers of the β phase with layer normals orthogonal to $\mathbf{E}$. We recall (1.2) and write it in the equivalent form

$$\begin{aligned} \alpha \Delta T &= 0 \quad \text{in the } \alpha \text{ phase,} \\ \beta \Delta T &= 0 \quad \text{in the } \beta \text{ phase, and} \\ \beta \nabla T \cdot \mathbf{n} &= \alpha \nabla T \cdot \mathbf{n} \quad \text{on the layer interface.} \end{aligned} \tag{4.33}$$

When $\mathbf{n}$ is perpendicular to $\mathbf{E}$ we easily see that the affine function $T = \mathbf{E} \cdot \mathbf{x}$ is a solution of (4.33) and optimality follows. To prove Theorem 1.3 we first suppose that that $1-\theta \leq 1/(\lambda-1)$ and show that the upper bound $\overline{\mathrm{B}}$ is saturated by a laminate of rank one. We refer to formulas (2.7), (2.9), and (2.10) and choose $\mathbf{n}^1$ parallel to $\mathbf{E}$ and set $\theta_1 = 1-\theta$ and $\theta_2 = 0$ to obtain

$$\lim_{\varepsilon \to 0} \|\nabla T_L^\varepsilon\|_2^2 = \overline{\mathrm{B}}. \tag{4.34}$$

Last we suppose that $1-\theta \geq 1/(\lambda-1)$. Here we choose $\mathbf{n}^1$ parallel to $\mathbf{E}$ and $\mathbf{n}^2$ orthogonal to $\mathbf{E}$ and choose $\theta_1 = \frac{1}{1+\theta(\lambda-1)}$ and $\theta_2 = 1-\theta-(\frac{1}{\lambda-1})$. This choice gives $\theta_1(1-\theta_2) = 1/(\lambda-1), (1-\theta_1)(1-\theta_2) = \theta$, and

$$\lim_{\varepsilon \to 0} \|\nabla T_L^\varepsilon\|_2^2 = \overline{\mathrm{B}} \tag{4.35}$$

follows from (2.7), (2.9), and (2.10).

REFERENCES

[1] J. Ball and R. James, *Fine phase mixtures as minimizers of energy*, Arch. Ration. Mech. Anal., 100 (1987), pp. 13–52.

[2] M. Briane, *Correctors for the homogenization of a laminate*, Adv. Math. Sci. Appl., 4 (1994), pp. 357–379.

[3] B. Dacorogna, *Direct Methods in the Calculus of Variations*, Springer-Verlag, Berlin, New York, 1989.
[4] I. Ekeland and R. Temam, *Convex Analysis and Variational Problems*, North-Holland, Amsterdam, 1976.
[5] K. Golden and G. Papanicolaou, *Bounds for effective parameters of heterogeneous media by analytic continuation*, Comm. Math. Phys., 90 (1983), pp. 473–491.
[6] R. V. Kohn and G. Strang, *Optimal design and relaxation of variational problems* I, II, *and* III, Comm. Pure Appl. Math., 39 (1986), pp. 113–137, 139–182, and 353–377.
[7] K. A. Lurie and A. V. Cherkaev, *Exact estimates of conductivity of composites formed by two isotropically conducting media taken in prescribed proportion*, Proc. Roy. Soc. Edinburgh Sect. A, 99 (1984), pp. 71–87.
[8] F. Murat and L. Tartar, *H-convergence*, in Topics in the Mathematical Modelling of Composite Materials, A. Cherkaev and R. V. Kohn, eds., Birkhäuser, Boston, 1997, pp. 21–43.
[9] F. Murat and L. Tartar, *Calcul des variations et homogénéisation,* in Les Méthodes de l′Homogénéisation: Théorie et Applications en Physique, Collect. Dir. Etudes et Rech. Elec. France 57, Eyrolles, Paris, 1985, pp. 319–369.
[10] S. Spagnolo, *Convergence in energy for elliptic operators*, in Proceedings of the Third Symposium on Numerical Solutions of Partial Differential Equations, B. Hubbard, ed., Academic Press, New York, 1976, pp. 496–498.
[11] L. Tartar, *Cours Peccot au Collège de France*, 1977.

SIAM J. MATH. ANAL.
Vol. 32, No. 5, pp. 1094–1116

STABILITY CONDITIONS FOR PATTERNS OF NONINTERACTING LARGE SHOCK WAVES*

MARTA LEWICKA†

Abstract. In this paper we study different conditions whose presence is required for

A. the admissibility and stability of large shocks present in solutions of a strictly hyperbolic $n \times n$ system of conservation laws in one space dimension

$$u_t + f(u)_x = 0,$$

B. the solvability and L^1 well posedness of the Cauchy problem for the above equation, near solutions containing large and stable, but noninteracting shock waves.

We compare the corresponding conditions of type A and B appearing in the literature; in particular, we show that the finiteness and stability conditions used in our most recent works generalize and/or unify these conditions in appropriate ways.

Key words. conservation laws, shock waves, stability conditions, large BV data

AMS subject classifications. 35L65, 35L45

PII. S0036141000367503

1. Introduction.

Consider the Cauchy problem for an $n \times n$ system of conservation laws in one space dimension:

$$u_t + f(u)_x = 0, \tag{1.1}$$

$$u(0, \cdot) = \bar{u}. \tag{1.2}$$

In the study of local existence and stability of solutions to (1.1), (1.2), due to the finite speed of propagation one is led to consider the special case where the initial data $\bar{u}$ is a small perturbation of a Riemann data:

$$\bar{u}(x) = \begin{cases} u^-, & x < 0, \\ u^+, & x > 0. \end{cases} \tag{1.3}$$

In this case, several results in the literature have shown that existence and stability of solutions can be obtained under a suitable linearized stability condition for the solutions of (1.1), (1.2), (1.3). (For a general theory of conservation laws in one space dimension, cf. [B], [D], [Sm].)

The main purpose of this paper is to compare the various assumptions of this kind and to prove their equivalence. We shall restrict ourselves to the case where the solution of (1.1), (1.2), (1.3) consists of $m + 1$ constant states, $m \in \{2, \dots, n\}$, separated by (possibly large) admissible shocks, say, in the characteristic families

*Received by the editors February 7, 2000; accepted for publication (in revised form) October 17, 2000; published electronically February 21, 2001. This research was partially supported by the European TMR Network on Hyperbolic Conservation Laws ERBFMRXCT960033.
http://www.siam.org/journals/sima/32-5/36750.html

†SISSA, via Beirut 2–4, 34014 Trieste, Italy (lewicka@sissa.it). Current address: Max Planck Institute for Mathematics in the Sciences (MIS), Inselstr. 22-26, 04103 Leipzig, Germany (lewicka@mis.mpg.de).

$i_1 < \cdots < i_m$. Calling these intermediate states $u_0^0 = u^-, u_0^1, u_0^2, \dots, u_0^m = u^+$, and Λ^q the speed of the i_q shock, the linearized system has the form

$$v_t + Df(u_0^q) \cdot v_x = 0, \qquad x/t \in (\Lambda^q, \Lambda^{q+1}). \tag{1.4}$$

Along shock lines we have the boundary equations obtained by linearizing the Rankine–Hugoniot equations that yield the linear dependence of the strengths of the outgoing waves on the components of the incoming wave vector interacting with the i_q large shock under consideration:

$$\epsilon_k^{out} = \sum_{\substack{s:1\dots n \\ incoming}} W_q^{k,s} \cdot \epsilon_s^{in} \tag{1.5}$$

(see Figure 1.1).

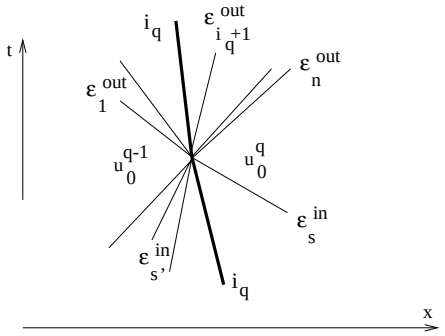


FIG. 1.1.

As we have mentioned, under some classical assumptions on the flux f in (1.1), which are recalled below, a variety of results concerning the (global) existence and uniqueness of admissible solutions to (1.1), (1.2) and their L^1 stability have been recently established [BC], [BM], [Le], [LT], [Scho], [W].

In all of these works, the conditions of two different natures are necessarily introduced:

A. conditions yielding the admissibility and stability of each of the large shocks in the reference solution of (1.1), (1.2), (1.3),

B. conditions guaranteeing the BV *stability* of the linearized system (1.4) [Scho], [BM], [Le], [W]. In [Scho], it is proved that they imply the local existence of solutions to the Cauchy problem, for data $\bar{u}$ suitably close to (1.3), and conditions providing the L^1 *stability* of the system (1.4) [BM], [BC], [Le]. It was proved that these in turn imply the L^1 stability of the nonlinear system (1.1), on a domain $\mathcal{D}$ of small BV perturbations of the data (1.3).

Our paper is organized as follows. In section 2 we focus on the conditions of type A, in particular, the well-known Majda stability condition [M].

Section 3 discusses different conditions of type B. In [Le], [LT], the stability conditions are formulated in terms of the existence of a suitable family of weights $w_s^q > 0$ such that the corresponding BV or L^1 norm of any solution of the linearized system (1.4) is nonincreasing in time. The main result of section 3 (Theorem 3.2) will show that the Schochet BV stability assumptions [Scho] are equivalent to BV stability assumptions in [Le]. Also, the L^1 stability condition in [BM], [Le], will appear to imply the mentioned BV stability (Theorem 3.1).

In the last section we treat the case of systems of $n = 2$ equations, with the presence of $m = 2$ large shocks and deal with the corresponding conditions introduced in [BC], [W], [LT].

We end this section by recalling the setting of the Cauchy problem (1.1), (1.2) (compare [Le]). In the n-dimensional state space $m+1$ distinct states $\{u_0^q\}_{q=0}^m$ are fixed, with their corresponding open disjoint neighborhoods $\{\Omega^q\}_{q=0}^m$ such that

- $f : \Omega \longrightarrow \mathbf{R}^n$ is smooth and defined on $\Omega = \bigcup_{q=0}^{m} \Omega^q \subset \mathbf{R}^n$.
- f is strictly hyperbolic in Ω, that is, at each point $u \in \Omega$, the matrix $Df(u)$ has n real and simple eigenvalues $\lambda_1(u) < \cdots < \lambda_n(u)$.
- Each characteristic field of (1.1) is either linearly degenerate or genuinely nonlinear, that is, with a basis $\{r_k(u)\}_{k=1}^n$ of corresponding right eigenvectors of $Df(u)$, $Df(u)r_k(u) = \lambda_k(u)r_k(u)$, each of the n directional derivatives $r_k \nabla \lambda_k$ vanishes either identically or nowhere.

The solution to (1.1), (1.2) with the initial data

$$\bar{u}(x) = \begin{cases} u_0^0, & x < 0, \\ u_0^m, & x > 0 \end{cases} \tag{1.6}$$

is given by m shocks (u_0^{q-1}, u_0^q), $q : 1 \dots m$, belonging to respective characteristic families i_q and travelling with respective speeds Λ^q:

$$u(t,x) = \begin{cases} u_0^0, & x < \Lambda^1 t, \\ u_0^q, & \Lambda^q t < x < \Lambda^{q+1} t, \quad q : 1 \dots m-1, \\ u_0^m, & x > \Lambda^m t, \end{cases} \tag{1.7}$$

as in Figure 1.2.

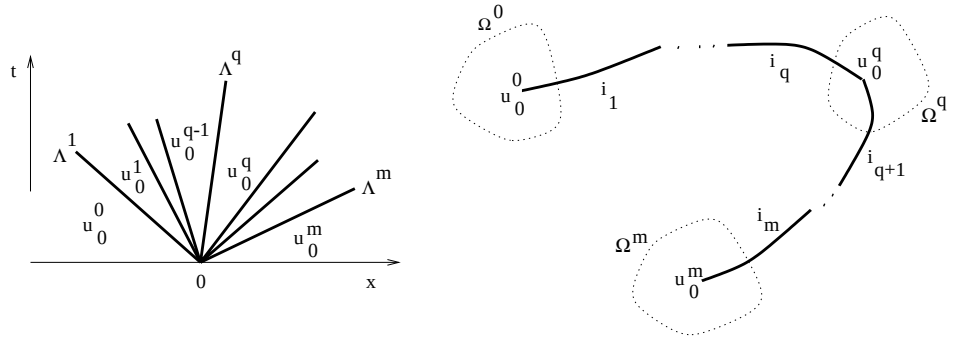


FIG. 1.2.

2. Stability of large shocks revisited. In this section we discuss the conditions of type A. Since every shock (u_0^{q-1}, u_0^q) has to be treated separately, it is not restrictive to assume that $m = 1$ and simplify the notation $u_0^0 = u^-, u_0^1 = u^+, \Omega^0 = \Omega^-, \Omega^1 = \Omega^m = \Omega^+, i_1 = i, \Lambda^1 = \Lambda$.

In this setting, for (1.7) to be a distributional solution of (1.1), (1.2), (1.3), the Rankine–Hugoniot conditions must be satisfied:

$$f(u^-) - f(u^+) = \Lambda(u^- - u^+). \tag{2.1}$$

Second, our i-shock is assumed to be compressive in the sense of Lax [L], that is,

$$\lambda_i(u^-) > \Lambda > \lambda_i(u^+). \tag{2.2}$$

Finally, in order to treat the Cauchy problem (1.1), (1.2), with $\bar{u}$ in (1.2) being a perturbation of (1.3), one must guarantee the so-called stability of the shock (u^-, u^+). This condition, introduced and justified in [LT], [Le], [BC] is the following:

(2.3)
There exists a smooth function $\Psi : \Omega^- \times \Omega^+ \longrightarrow \mathbf{R}^{n-1}$ such that

(i) $\Psi(u^0, u^1) = 0$ iff the states u^0 and u^1 can be connected by a (large) shock of the ith characteristic family, with the speed $\Lambda(u^0, u^1)$. The Rankine–Hugoniot condition holds: $f(u^0) - f(u^1) = \Lambda(u^0, u^1)(u^0 - u^1)$. In particular, $\Psi(u^-, u^+) = 0$ and $\Lambda(u^-, u^+) = \Lambda$.

(ii)
$$\operatorname{rank} \frac{\partial \Psi}{\partial u^0}(u^-, u^+) = \operatorname{rank} \frac{\partial \Psi}{\partial u^1}(u^-, u^+) = n - 1.$$

(iii) The $n - 1$ vectors
$$\left\{ \frac{\partial \Psi}{\partial u^0}(u^-, u^+) \cdot r_k(u^-) \right\}_{k=1}^{i-1} \cup \left\{ \frac{\partial \Psi}{\partial u^1}(u^-, u^+) \cdot r_k(u^+) \right\}_{k=i+1}^{n}$$
are linearly independent.

Under these hypotheses one can see that if only the sets Ω^-, Ω^+ are small enough, then any Riemann problem $(u^0, u^1) \in \Omega^- \times \Omega^+$ for (1.1) has a unique self-similar solution composed of n shocks or rarefaction waves. The ith wave is a large i compressive Lax shock, connecting some states in the domains Ω^- and Ω^+.

In [Scho], the stability of the large shock (u^-, u^+) satisfying (2.1), (2.2) is understood in the classical sense of Majda:

(2.4)
The n vectors
$$r_1(u^-), \dots, r_{i-1}(u^-), u^- - u^+, r_{i+1}(u^+), \dots, r_n(u^+)$$
are linearly independent.

Obviously for weak shocks (2.4) is always satisfied, and equivalent to (2.3)(iii). The main result of this section discusses this same situation in the general case.

THEOREM 2.1. *Let (u^-, u^+) be a Rankine–Hugoniot shock such that its speed Λ in* (2.1) *is not an eigenvalue of $Df(u^-)$ nor of $Df(u^+)$. Then the conditions* (2.3) *and* (2.4) *are equivalent.*

The proof of Theorem 2.1 relies on the construction of a particular function Ψ_0, whose zero level set consists of those pairs of states $(u^0, u^1) \in \Omega^- \times \Omega^+$ that can be connected by an admissible i-shock as in (2.3)(i).

We define Ψ_0 as follows:

$$\Psi_0(u^0, u^1) = \left\{ \left\langle f(u^1) - f(u^0),\ V_k(u^1 - u^0) \right\rangle \right\}_{k=1}^{n-1}, \tag{2.5}$$

where V_k are any smooth functions defined on a neighborhood of the vector $u_0 = u^+ - u^- \neq 0$ with values in $\mathbf{R}^n$, and such that for every u the space

$$\text{span}\{V_1(u), \dots, V_{n-1}(u)\}$$

is the orthogonal complement of the vector u.

LEMMA 2.2. *$\{V_k\}_{k=1}^{n-1}$ can be taken so that*

$$V_k(u_0) = -[DV_k(u_0)]^T \cdot u_0 \quad \forall k : 1 \dots n-1. \tag{2.6}$$

Proof. By $e_1, \dots, e_n$ we denote the standard Euclidean base of $\mathbf{R}^n$.

For u close to e_n define the vectors $\{\widetilde{V}_k(u)\}_{k=1}^{n-1}$ applying the Gramm–Schmidt orthogonalization process to n linearly independent vectors $u, e_1, \dots, e_{n-1}$. Namely, set

$$\begin{aligned} \widetilde{V}_1(u) &= e_1 - \langle e_1, u\rangle \cdot \tfrac{u}{|u|^2}, \\ \widetilde{V}_k(u) &= e_k - \left[\langle e_k, u\rangle \cdot \tfrac{u}{|u|^2} + \sum_{s=1}^{k-1} \langle e_k, \widetilde{V}_s(u)\rangle \cdot \widetilde{V}_s(u)\right] \quad \forall k : 2 \dots n-1. \end{aligned} \tag{2.7}$$

Note that

$$\widetilde{V}_k(e_n) = e_n \quad \forall k : 1 \dots n-1 \tag{2.8}$$

and

- $\langle \widetilde{V}_k(u), u\rangle = 0 \quad \forall k : 1 \dots n-1$,
- $\{\widetilde{V}_k\}_{k=1}^{n-1}$ are smooth functions of u.

Thus, $\text{span}\{\widetilde{V}_1(u), \dots, \widetilde{V}_{n-1}(u)\}$ always complements orthogonally the vector u.

Moreover, using (2.8) and the fact that $\widetilde{V}_k \in \text{span}(e_1, \dots, e_k, u)$, by the explicit formulas (2.7) one proves inductively that

$$D\widetilde{V}_k(e_n) = [d_{sl}]_{s,l:1\dots n}, \quad d_{sl} = \begin{cases} -1 & \text{for } (s,l) = (n,k), \\ 0 & \text{otherwise.} \end{cases} \tag{2.9}$$

Now for u close to u_0 define

$$V_k(u) = A^{-1} \cdot \widetilde{V}_k(Au), \tag{2.10}$$

where A is an orthogonal transformation composed with an appropriate dilation such that $Au_0 = e_n$. Consequently

$$A^{-1} = |u_0|^2 A^T. \tag{2.11}$$

Obviously $\{V_k\}_{k=1}^{n-1}$ are smooth functions, and by the corresponding property of $\{\widetilde{V}_k\}_{k=1}^{n-1}$ they span the orthogonal complement of its argument vector.

By (2.10), (2.11), (2.9), and (2.8) we get

$$\begin{aligned} [DV_k(u_0)]^T \cdot u_0 =& A^T \cdot [D\widetilde{V}_k(e_n)]^T \cdot (A^T)^{-1} \cdot u_0 = A^{-1} \cdot [D\widetilde{V}_k(e_n)]^T \cdot Au_0 \\ =& -A^{-1} e_k = -A^{-1} \cdot \widetilde{V}_k(Au_0) = -V_k(u_0), \end{aligned}$$

which proves (2.6). ∎

Using the above lemma one finds a convenient formula for the derivatives of Ψ_0:

$$\frac{\partial \Psi_0}{\partial u^0}(u^-, u^+) = -V \cdot [Df(u^-) - \Lambda Id], \tag{2.12}$$

$$\frac{\partial \Psi_0}{\partial u^1}(u^-, u^+) = V \cdot [Df(u^+) - \Lambda Id], \tag{2.13}$$

where V is the $(n-1) \times n$ matrix, whose rows are the vectors $V_1(u_0), \dots, V_{n-1}(u_0)$. Note that since rank $V = n-1$, then Λ is neither an eigenvalue of $Df(u^-)$ nor $Df(u^+)$, which in view of (2.12), (2.13) implies

$$\operatorname{rank} \frac{\partial \Psi_0}{\partial u^0}(u^-, u^+) = \operatorname{rank} \frac{\partial \Psi_0}{\partial u^1}(u^-, u^+) = n-1. \tag{2.14}$$

Proof of Theorem 2.1.

Step 1. By (2.12), (2.13) we get

$$\frac{\partial \Psi_0}{\partial u^0}(u^-, u^+) \cdot r_k(u^-) = -(\lambda_k(u^-) - \Lambda) \cdot V \cdot r_k(u^-) \quad \forall k : 1 \dots i-1,$$

$$\frac{\partial \Psi_0}{\partial u^1}(u^-, u^+) \cdot r_k(u^+) = (\lambda_k(u^+) - \Lambda) \cdot V \cdot r_k(u^+) \quad \forall k : i+1 \dots n.$$

Since $\Lambda \notin \{\lambda_k(u^-)\}_{k=1}^{i-1} \cup \{\lambda_k(u^+)\}_{k=i+1}^{n}$ we see that the condition (2.3)(iii) for our function Ψ_0 is satisfied iff the vectors $\{V \cdot r_k(u^-)\}_{k=1}^{i-1} \cup \{V \cdot r_k(u^+)\}_{k=i+1}^{n}$ are linearly independent, which is in turn equivalent to Majda's condition (2.4), as $\ker\ V = \operatorname{span}(u_0)$. We have thus shown that (2.4) is equivalent to (2.3)(iii) for the function Ψ_0.

Recalling (2.14), one sees this way that (2.4) implies (2.3).

Step 2. Now we turn toward proving the converse implication. Let Ψ be any function satisfying (2.3). In particular, by (2.3)(ii), $\operatorname{rank}\ D\Psi(u^-, u^+)$ is maximal and equal to $n-1$. The same is true for $D\Psi_0(u^-, u^+)$, by (2.14), so

$$\operatorname{rank}\ D\Psi(u^-, u^+) = \operatorname{rank}\ D\Psi_0(u^-, u^+). \tag{2.15}$$

Another important remark is that

$$\ker\ D\Psi(u^-, u^+) = \ker\ D\Psi_0(u^-, u^+). \tag{2.16}$$

The spaces in (2.16) both coincide with the tangent space of the manifold $(\Psi_0)^{-1}(0)$ at point (u^-, u^+).

The following simple fact of linear algebra will be used in what follows.

LEMMA 2.3. *Let $A, B : \mathbf{R}^n \longrightarrow \mathbf{R}^s$ be two linear operators, $s < n$. Assume that* $\operatorname{rank}\ A = \operatorname{rank}\ B = s$ *and* $\ker\ A = \ker\ B$. *Then for any s vectors $v_1, \dots, v_s \in \mathbf{R}^n$ it holds that the vectors $\{Av_k\}_{k=1}^s$ are linearly independent iff $\{Bv_k\}_{k=1}^s$ are linearly independent.*

In view of (2.15), (2.16), we can apply Lemma 2.2 to the linear operators

$$D\Psi(u^-, u^+),\ D\Psi_0(u^-, u^+) : \mathbf{R}^{2n} \longrightarrow \mathbf{R}^{n-1}$$

and the following set of $n-1$ test vectors in $\mathbf{R}^{2n}$:

$$\{[r_k(u^-)^T, 0 \dots 0]^T\}_{k=1}^{i-1} \cup \{[0 \dots 0, r_k(u^+)^T]^T\}_{k=i+1}^{n}.$$

By (2.3)(iii) we receive that the same condition is satisfied by our function Ψ_0. This in turn, is equivalent to (2.4), as shown in Step 1. □

The proof of Theorem 2.1 shows that if the function Ψ as in (2.3) exists, then it can be replaced by the function Ψ_0, in this case necessarily enjoying the properties (2.3)(i)–(2.3)(iii).

3. *BV* and L^1 stability conditions compared. In this and the next sections we discuss different stability conditions of type B, used in [BC], [W], [Scho], and [Le]. Recall that these conditions guarantee the well posedness of the problem (1.1), (1.2) and the existence of the Lipschitz continuous semigroup of solutions, whose domain contains all the small $L^1 \cap BV$ perturbations of the initial data $\bar{u}$ in (1.6) (compare [Le]).

We show the equivalence of the Schochet BV stability condition (called in [Scho] the finiteness condition) with the BV stability condition used in [Le], as well as with the Wang BV stability condition [W], and the equivalence of L^1 stability condition from [BM], [Le] with the one introduced in [BC] for 2×2 systems.

Also (see Remark 3.8), we position our work to some of the results found in [LY].

We start by recalling the mentioned conditions.

3.1. *BV* stability condition. There exist positive weights $w_1^q, \dots, w_n^q$ (for every $q : 0 \dots m$) such that the following holds. Consider a small wave of a family $k \leq i_q$, hitting from the right the large initial i_q-shock (u_0^{q-1}, u_0^q), as in Figure 3.1. Then

$$\sum_{s=1}^{i_q-1} \frac{w_s^{q-1}}{w_k^q} \cdot \left| \frac{\partial}{\partial \epsilon_k^{in}} \, \epsilon_s^{out} \right| + \sum_{s=i_q+1}^{n} \frac{w_s^q}{w_k^q} \cdot \left| \frac{\partial}{\partial \epsilon_k^{in}} \, \epsilon_s^{out} \right| < 1 \tag{3.1}$$

at $\epsilon_1^{in} = \cdots = \epsilon_k^{in} = \cdots = \epsilon_n^{in} = 0$.

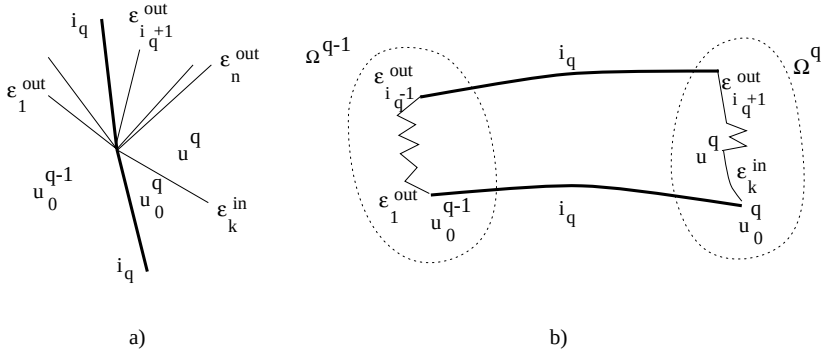


FIG. 3.1.

Symmetrically, in the case when a small k-wave with $k \geq i_q$ hits the shock (u_0^{q-1}, u_0^q) from the left (compare Figure 3.2), there holds

$$\sum_{s=1}^{i_q-1} \frac{w_s^{q-1}}{w_k^{q-1}} \cdot \left| \frac{\partial}{\partial \epsilon_k^{in}} \, \epsilon_s^{out} \right| + \sum_{s=i_q+1}^{n} \frac{w_s^q}{w_k^{q-1}} \cdot \left| \frac{\partial}{\partial \epsilon_k^{in}} \, \epsilon_s^{out} \right| < 1 \tag{3.2}$$

at $\epsilon_1^{in} = \cdots = \epsilon_k^{in} = \cdots = \epsilon_n^{in} = 0$.

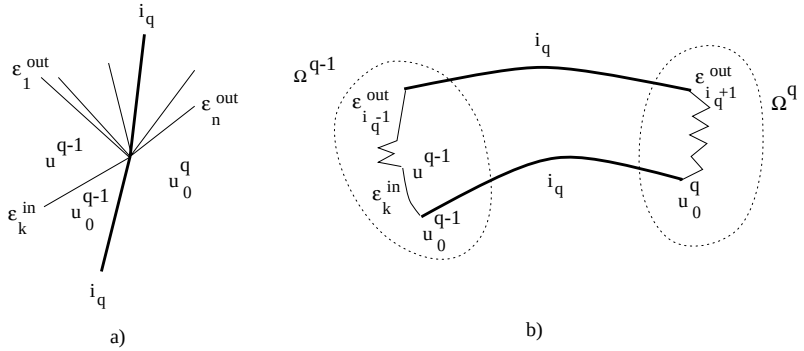


FIG. 3.2.

Regarding w_s^q as the weight given to an s-wave located in the region between the $q-1$ and the qth large shock, conditions (3.1), (3.2) simply say that every time a small wave hits a large shock, the total weighted strength of the outgoing small waves is smaller than the weighted strength of the incoming wave.

3.2. L^1 stability condition [Le], [BM]. There exist positive weights $w_1^q, \dots, w_n^q$ (for every $q : 0 \dots m$) such that in the setting of Figure 3.1

$$\begin{aligned}\sum_{s=1}^{i_q-1} \frac{w_s^{q-1}}{w_k^q} \cdot \left| \frac{\partial}{\partial \epsilon_k^{in}} \left(\frac{\epsilon_s^{out} \cdot (\lambda_s^{out} - \Lambda^q)}{(\lambda_k^{in} - \Lambda^q)} \right) \right| \\ + \sum_{s=i_q+1}^{n} \frac{w_s^q}{w_k^q} \cdot \left| \frac{\partial}{\partial \epsilon_k^{in}} \left(\frac{\epsilon_s^{out} \cdot (\lambda_s^{out} - \Lambda^q)}{(\lambda_k^{in} - \Lambda^q)} \right) \right| < 1\end{aligned} \tag{3.3}$$

at $\epsilon_1^{in} = \cdots = \epsilon_k^{in} = \cdots = \epsilon_n^{in} = 0$, while in the setting of Figure 3.2

$$\begin{aligned}\sum_{s=1}^{i_q-1} \frac{w_s^{q-1}}{w_k^{q-1}} \cdot \left| \frac{\partial}{\partial \epsilon_k^{in}} \left(\frac{\epsilon_s^{out} \cdot (\lambda_s^{out} - \Lambda^q)}{(\lambda_k^{in} - \Lambda^q)} \right) \right| \\ + \sum_{s=i_q+1}^{n} \frac{w_s^q}{w_k^{q-1}} \cdot \left| \frac{\partial}{\partial \epsilon_k^{in}} \left(\frac{\epsilon_s^{out} \cdot (\lambda_s^{out} - \Lambda^q)}{(\lambda_k^{in} - \Lambda^q)} \right) \right| < 1\end{aligned} \tag{3.4}$$

at $\epsilon_1^{in} = \cdots = \epsilon_k^{in} = \cdots = \epsilon_n^{in} = 0$.

Note that since the weights $\{w_i^0\}_{i=1}^n$ and $\{w_i^m\}_{i=1}^n$ appear only in one inequality (3.1) or (3.2), then the corresponding BV stability estimates for the leftmost large shock (u_0^0, u_0^1) and the rightmost (u_0^{m-1}, u_0^m) may take the following, simplified form:

$$\sum_{s=i_2}^{n} \frac{w_s^1}{w_k^1} \cdot \left| \frac{\partial}{\partial \epsilon_k^{in}} \, \epsilon_s^{out} \right| < 1 \tag{3.1a}$$

for all small waves of families $k \le i_1$, hitting the first shock i_1 from the right, and

$$\sum_{s=1}^{i_{m-1}} \frac{w_s^{m-1}}{w_k^{m-1}} \cdot \left| \frac{\partial}{\partial \epsilon_k^{in}} \, \epsilon_s^{out} \right| < 1 \tag{3.2a}$$

for all small waves of families $k \ge i_m$, hitting the last shock i_m from the left.

The analogous simplifications may be easily done for the L^1 stability estimates (3.3) and (3.4).

Also, for $q \notin \{1, m\}$, (3.1) and (3.2) can be rewritten as follows:

$$\sum_{s=1}^{i_{q-1}} \frac{w_s^{q-1}}{w_k^q} \cdot \left| \frac{\partial}{\partial \epsilon_k^{in}} \, \epsilon_s^{out} \right| + \sum_{s=i_{q+1}}^{n} \frac{w_s^q}{w_k^q} \cdot \left| \frac{\partial}{\partial \epsilon_k^{in}} \, \epsilon_s^{out} \right| < 1, \tag{3.1a}$$

$$\sum_{s=1}^{i_{q-1}} \frac{w_s^{q-1}}{w_k^{q-1}} \cdot \left| \frac{\partial}{\partial \epsilon_k^{in}} \, \epsilon_s^{out} \right| + \sum_{s=i_{q+1}}^{n} \frac{w_s^q}{w_k^{q-1}} \cdot \left| \frac{\partial}{\partial \epsilon_k^{in}} \, \epsilon_s^{out} \right| < 1. \tag{3.2a}$$

Analogously, the L^1 stability condition (3.3) and (3.4) for $q \notin \{1, m\}$ may be formulated with the correspondingly changed summation ranges.

THEOREM 3.1. *The L^1 stability condition* (3.3), (3.4) *implies the BV stability condition* (3.1), (3.2).

Proof. In view of the preceding remarks, assume that the L^1 stability condition (3.3a) and (3.4a) holds, with weights $\{w_s^q\}$. For $q : 1 \dots m-1$ and $s : 1 \dots n$ define

$$\widetilde{w}_s^q = \left| \lambda_s(u_0^q) - \Lambda^{q+1} \right| \cdot w_s^q.$$

We will show that the BV stability condition (3.1), (3.2) is satisfied for all $q : 1 \dots m$.

Indeed, to prove (3.1), compute

$$\begin{aligned}
&\sum_{s=1}^{i_{q-1}} \frac{\widetilde{w}_s^{q-1}}{\widetilde{w}_k^q} \cdot \left| \frac{\partial}{\partial \epsilon_k^{in}} \, \epsilon_s^{out} \right| + \sum_{s=i_{q+1}}^{n} \frac{\widetilde{w}_s^q}{\widetilde{w}_k^q} \cdot \left| \frac{\partial}{\partial \epsilon_k^{in}} \, \epsilon_s^{out} \right| \\
&\qquad = \sum_{s=1}^{i_{q-1}} \frac{w_s^{q-1}}{w_k^q} \cdot \left| \frac{\partial}{\partial \epsilon_k^{in}} \, \epsilon_s^{out} \right| \cdot \frac{|\lambda_s(u_0^{q-1}) - \Lambda^q|}{|\lambda_k(u_0^q) - \Lambda^{q+1}|} \\
&\qquad\quad + \sum_{s=i_{q+1}}^{n} \frac{w_s^q}{w_k^q} \cdot \left| \frac{\partial}{\partial \epsilon_k^{in}} \, \epsilon_s^{out} \right| \cdot \frac{|\lambda_s(u_0^q) - \Lambda^{q+1}|}{|\lambda_k(u_0^q) - \Lambda^{q+1}|} < 1
\end{aligned}$$

by (3.3a) and the following easily received inequalities:

$$\begin{aligned}
&\left| \lambda_k(u_0^q) - \Lambda^{q+1} \right| > |\lambda_k(u_0^q) - \Lambda^q| \quad \forall k \le i_q, \\
&\left| \lambda_k(u_0^q) - \Lambda^{q+1} \right| < |\lambda_k(u_0^q) - \Lambda^q| \quad \forall k \ge i_{q+1}.
\end{aligned}$$

The estimate (3.2) is justified in a similar way. ∎

3.3. The Schochet BV stability condition [Scho]. In connection with (3.1) and (3.2), for every $q : 1 \dots m$ define four nonnegative matrices, expressing the strengths of outgoing waves in terms of the strengths of the incoming small waves,

interacting with the large initial i_q-shock:

- interaction from the right, waves outgoing to the right

$$M_q^{rr} = [a_{sk}^q], \quad s : i_{q+1} \dots n, \ k : 1 \dots i_q,$$

- interaction from the right, waves outgoing to the left

$$M_q^{rl} = [a_{sk}^q], \quad s : 1 \dots i_{q-1}, \ k : 1 \dots i_q,$$

(3.5)

- interaction from the left, waves outgoing to the right

$$M_q^{lr} = [a_{sk}^q], \quad s : i_{q+1} \dots n, \ k : i_q \dots n,$$

- interaction from the left, waves outgoing to the left

$$M_q^{ll} = [a_{sk}^q], \quad s : 1 \dots i_{q-1}, \ k : i_q \dots n.$$

In all of the above definitions

$$a_{sk}^q = \left| \frac{\partial \epsilon_s^{out}}{\partial \epsilon_k^{in}} \right|$$

at $\epsilon_1^{in} = \cdots = \epsilon_k^{in} = \cdots = \epsilon_n^{in} = 0$.

Note that in (3.5) the range of s (indexing the outgoing small waves) depends on the neighboring large shock (of the family i_{q-1} or i_{q+1}). Indeed, it is relevant to keep track of only these newborn waves that in the future may possibly interact with large shocks, thus changing the global wave pattern.

Keeping the above comment in mind, we also remark that the notation for the matrices $M_1^{rl}, M_1^{lr}, M_1^{ll}, M_m^{rr}, M_m^{lr}, M_m^{rl}$ is ambiguous, however, in view of what we have said, the precise form of these matrices is irrelevant in the following analysis.

Consider the first pair of large shocks (u_0^0, u_0^1) and (u_0^1, u_0^2) and a tuple $\gamma = [\gamma_k]_{k:i_2 \dots n}$ of small waves travelling in the region between these shocks, and approaching the second one. By interaction of γ with (u_0^1, u_0^2), then, interaction of the newborn "reflected" waves with (u_0^0, u_0^1) and so on, further waves travelling in the region between the two large shocks are produced. Call

$$R^1 = M_1^{rr}. \tag{3.6}$$

The total strength of such waves, belonging to the characteristic families $k \geq i_2$, is then seen to be

$$\left[Id + R^1 M_2^{ll} + \left(R^1 M_2^{ll} \right)^2 + \cdots \right] |\gamma| = \left(Id - R^1 M_2^{ll} \right)^{-1} |\gamma| \doteq P^{1-2} |\gamma|$$

(where $|\gamma| = [|\gamma_k|]_{k:i_2 \dots n}$), provided that the first finiteness requirement

$$\text{all eigenvalues of } R^1 \cdot M_2^{ll} \text{ are } < 1 \text{ in absolute value} \tag{3.7}$$

is satisfied.

Now, view the pair of the first two large shocks as a single entity. The reflection matrix R^{1-2}, expressing the strengths of the outgoing small waves of families $k \geq i_3$, exiting the region between the first and the second large waves to the right of the

latter one, in terms of the incoming waves of the families $k \leq i_2$, possibly interacting with the $(i_1 - i_2)$ couple of large shocks from the right, has the form

$$R^{1-2} = M_2^{rr} + M_2^{lr} P^{1-2} R^1 M_2^{rl}.$$

The natural finiteness requirement for the triple $(i_1 - i_2 - i_3)$ of large shocks, analogous to (3.7) is then

$$\text{all eigenvalues of } R^{1-2} \cdot M_3^{ll} \text{ are } < 1 \text{ in absolute value.}$$

Proceeding in the same manner and viewing any fixed combination $(i_1 - \cdots - i_q)$ of consecutive large shocks as a single entity, influencing its succeeding large wave i_{q+1}, we obtain the following $(m-1)$ assertions that constitute the announced Schochet BV stability condition:

$$\begin{aligned} &\text{spRad}(F^{1-2}) < 1, \\ &\text{spRad}(F^{1-2-3}) < 1, \\ &\quad \vdots \\ &\text{spRad}(F^{1-\cdots-m}) < 1 \end{aligned} \tag{3.8}$$

(spRad stands here for the spectral radius of the reference matrix). The finiteness matrices F are defined inductively, together with the corresponding reflection and production matrices R, P, by recalling (3.6) and setting

$$F^{1-\cdots-q} \doteq R^{1-\cdots-(q-1)} \cdot M_q^{ll} \qquad \text{for } q : 2 \dots m, \tag{3.9}$$

$$P^{1-\cdots-q} \doteq \left(Id - F^{1-\cdots-q}\right)^{-1} \qquad \text{for } q : 2 \dots m, \tag{3.10}$$

$$R^{1-\cdots-q} \doteq M_q^{rr} + M_q^{lr} P^{1-\cdots-q} R^{1-\cdots-(q-1)} M_q^{rl} \qquad \text{for } q : 2 \dots m-1. \tag{3.11}$$

3.4. *BV* stability condition. The main theorem of this subsection is the following.

THEOREM 3.2. *The BV stability condition* (3.1), (3.2) *is equivalent to the Schochet BV stability condition* (3.8).

To prove Theorem 3.2, we need two abstract results on matrix theory.

LEMMA 3.3. *Let $Q = [q_{sk}]_{s,k:1\dots n}$ be an $n \times n$ matrix with nonnegative entries: $q_{sk} \geq 0$. The following conditions are equivalent:*

(i) $\text{spRad}(Q) < 1$.

(ii) *There exists a diagonal matrix $W = \text{diag}(w_1, \dots, w_n)$ with positive diagonal entries $w_s > 0$ such that $\| WQW^{-1} \|_1 < 1$.*
Here the norm of an $n \times n$ matrix $P = [p_{sk}]_{s,k:1\dots n}$ is defined by

$$\| P \|_1 \ = \max_{k:1\dots n} \sum_{s=1}^{n} |p_{sk}|.$$

The above lemma, which came up independently in the investigations leading to this paper, follows also from the results in [LY, Theorem 1 in Appendix 1]; thus for brevity we omit its proof.

LEMMA 3.4. *Let A, B be two $n \times n$ matrices with nonnegative entries:*

$$A = [a_{sk}]_{s,k:1\dots n}, \quad B = [b_{sk}]_{s,k:1\dots n}.$$

Assume that there exist two sets of indices $col, ver \subset \{1 \dots n\}$ *with the properties*

- $col \cap ver = \emptyset$,
- $\forall k \notin col\ \forall s : 1 \dots n, \quad a_{sk} = b_{ks} = 0$,
- $\forall s \notin ver\ \forall k : 1 \dots n, \quad a_{sk} = b_{ks} = 0$.

Then the following two statements are equivalent:

(i) *There exists* $W = \mathrm{diag}(w_1, \dots, w_n)$ *with all* $w_k > 0$ *such that* $\| WAW^{-1} \|_1 < 1$ *and* $\| WBW^{-1} \|_1 < 1$.

(ii) *There exists* $W = \mathrm{diag}(w_1, \dots, w_n)$ *with all* $w_k > 0$ *such that* $\| WABW^{-1} \|_1 < 1$.

The matrix norm $\| \cdot \|_1$ *is defined as in Lemma* 3.3.

Proof. (i) $\Rightarrow$ (ii). This implication is an obvious consequence of the fact that $\| \cdot \|_1$ is a matrix norm.

(ii) $\Rightarrow$ (i). Since $WABW^{-1} = (WAW^{-1})(WBW^{-1})$, we may without loss of generality assume that $\| AB \|_1 < 1$ and prove the existence of a diagonal matrix W satisfying (i). By (ii) we have

$$\sum_{s \in col} \left[b_{sk} \cdot \sum_{r \in ver} a_{rs} \right] < 1 \quad \forall k \in ver.$$

For a fixed $\epsilon > 0$ define

$$w_k = \begin{cases} \displaystyle\sum_{s \in ver} a_{sk} + \epsilon & \text{for } k \in col, \\ 1 & \text{otherwise.} \end{cases}$$

Then

$$\sum_{s \in ver} w_s a_{sk} = \sum_{s \in ver} a_{sk} < w_k \quad \forall k \in col,$$

$$\sum_{s \in col} w_s b_{sk} = \sum_{s \in col} \left(\sum_{r \in ver} a_{rs} \right) b_{sk} + \sum_{s \in col} \epsilon b_{sk} < 1 = w_k \quad \forall k \in ver,$$

provided that ϵ is small enough.

We have thus proved that $\| WAW^{-1} \|_1 < 1$ and $\| WBW^{-1} \|_1 < 1$. ∎

For every matrix $M_q^{xy},\ x, y \in \{l, r\}$, define the corresponding square $n \times n$ matrix $\widetilde{M}_q^{xy}$ by completing all the "missing" entries with zeros. For example, in view of (3.5)

$$\widetilde{M}_1^{rr} = [\widetilde{a}_{sk}]_{s,k:1\dots n}, \qquad \widetilde{a}_{sk} = \begin{cases} a_{sk} & \text{for } s : i_2 \dots n,\ k : 1 \dots i_1, \\ 0 & \text{otherwise.} \end{cases}$$

The next lemma shows some possible reformulations of our BV stability condition (3.1), (3.2).

LEMMA 3.5. *The following conditions are equivalent to the* BV *stability condition* (3.1), (3.2) :

(i) *There exist $m-1$ diagonal matrices $\{W^q\}_{q=1}^{m-1}$ with positive diagonal entries such that*

(3.12)
$$\| W^1 \widetilde{M}_1^{rr} (W^1)^{-1} \|_1 < 1,$$

(3.13)
$$\begin{array}{l} \| W^{q-1} \widetilde{M}_q^{ll} (W^{q-1})^{-1} + W^q \widetilde{M}_q^{lr} (W^{q-1})^{-1} \|_1 < 1 \\ \| W^q \widetilde{M}_q^{rr} (W^q)^{-1} + W^{q-1} \widetilde{M}_q^{rl} (W^q)^{-1} \|_1 < 1 \end{array} \qquad \forall q : 2 \dots m-1,$$

(3.14)
$$\| W^{m-1} \widetilde{M}_m^{ll} (W^{m-1})^{-1} \|_1 < 1.$$

(ii) *Define two block square matrices of the dimension $(m-1)\cdot n$:*

$$Odd_m = \begin{bmatrix} \widetilde{M}_1^{rr} & 0 & \dots & \dots & 0 \\ 0 & \widetilde{M}_3^{ll} & \widetilde{M}_3^{rl} & 0 & \vdots \\ \vdots & \widetilde{M}_3^{lr} & \widetilde{M}_3^{rr} & 0 & \\ \vdots & 0 & 0 & \widetilde{M}_5^{ll} & \\ 0 & \dots & & & \ddots \end{bmatrix},$$

$$Even_m = \begin{bmatrix} \widetilde{M}_2^{ll} & \widetilde{M}_2^{rl} & 0 & \dots & 0 \\ \widetilde{M}_2^{lr} & \widetilde{M}_2^{rr} & 0 & \dots & \\ 0 & 0 & \widetilde{M}_4^{ll} & \widetilde{M}_4^{rl} & \\ \vdots & \vdots & \widetilde{M}_4^{lr} & \widetilde{M}_4^{rr} & \\ 0 & & & & \ddots \end{bmatrix}.$$

Then

(3.15)
$$\mathrm{spRad}(Odd_m \cdot Even_m) < 1.$$

Proof. The condition (i) is obviously equivalent to (3.1), (3.2) if we define $W^q = \mathrm{diag}(w_1^q, \dots, w_n^q)$ for all $q : 1 \dots m-1$.

Note that (3.12), (3.13), (3.14) are equivalent to

(3.16)
$$\| W \cdot Odd_m \cdot W^{-1} \| < 1, \qquad \| W \cdot Even_m \cdot W^{-1} \| < 1,$$

where W is the block diagonal matrix of the dimension $(m-1)\cdot n$ given by

$$W = \mathrm{diag}(W^1, \dots, W^{m-1}).$$

By Lemma 3.3 and Lemma 3.4, (3.16) is in turn equivalent to (3.15), which proves (ii). ∎

Before we give the proof of Theorem 3.2, we need one more result of a technical nature.

LEMMA 3.6. *Let A, B be two $n \times n$ matrices with nonnegative entries such that $\| A + B \|_1 < 1$. Then $\| B \cdot (Id - A)^{-1} \|_1 < 1$.*

Proof. Note first that since $\| A \|_1 < 1$, then the matrix $Id - A$ is invertible and its inverse

$$(Id - A)^{-1} = Id + A + A^2 + \cdots$$

has nonnegative entries. From the assumption it follows moreover that

$$\sum_{i=1}^{n} [B]_{ik} < 1 - \sum_{i=1}^{n} [A]_{ik} = \sum_{i=1}^{n} [Id - A]_{ik},$$

for every $k : 1 \dots n$, and thus

$$\begin{aligned}
\sum_{i=1}^{n} [B \cdot (Id - A)^{-1}]_{ik} &= \sum_{s=1}^{n} \left(\sum_{i=1}^{n} [B]_{is} \right) \cdot [(Id - A)^{-1}]_{sk} \\
&< \sum_{s=1}^{n} \left(\sum_{i=1}^{n} [Id - A]_{is} \right) \cdot [(Id - A)^{-1}]_{sk} \\
&= \sum_{i=1}^{n} [(Id - A) \cdot (Id - A)^{-1}]_{ik} = 1,
\end{aligned}$$

for every $k : 1 \dots n$, which proves our lemma. □

Now we are ready to give the following proof.

Proof of Theorem 3.2.

Step 1. (3.1), (3.2) ⇒ (3.8). We use the equivalent form of the BV stability condition (3.1), (3.2) given in Lemma 3.5(i).

We first show that

$$\forall q : 1 \dots m-1 \quad \| W^q \cdot \widetilde{R}^{1-\cdots-q} \cdot (W^q)^{-1} \|_1 < 1. \tag{3.17}$$

We proceed by induction on q. For $q = 1$, (3.17) is equivalent to (3.12) in view of (3.6). For $q : 2 \dots m-1$, by (3.11) we have

$$\begin{aligned}
W^q \cdot \widetilde{R}^{1-\cdots-q} \cdot (W^q)^{-1} &= W^q \widetilde{M}_q^{rr} (W^q)^{-1} \\
&\quad + \left[W^q \widetilde{M}_q^{lr} \widetilde{P}^{1-\cdots-q} \widetilde{R}^{1-\cdots-(q-1)} (W^{q-1})^{-1} \right] \\
&\quad \cdot \left[W^{q-1} \widetilde{M}_q^{rl} (W^q)^{-1} \right].
\end{aligned}$$

The desired conclusion (3.17) will thus follow from the second inequality in (3.13) provided that

$$\| W^q \widetilde{M}_q^{lr} \widetilde{P}^{1-\cdots-q} \widetilde{R}^{1-\cdots-(q-1)} (W^{q-1})^{-1} \|_1 < 1. \tag{3.18}$$

Note that

$$
\begin{aligned}
&W^q\widetilde{M}_q^{lr}\widetilde{P}^{1-\cdots-q}\widetilde{R}^{1-\cdots-(q-1)}(W^{q-1})^{-1}\\
&\quad= W^q\widetilde{M}_q^{lr}\cdot\Big(Id-\widetilde{R}^{1-\cdots-(q-1)}\widetilde{M}_q^{ll}\Big)^{-1}\cdot\widetilde{R}^{1-\cdots-(q-1)}(W^{q-1})^{-1}\\
&\quad= \Big[W^q\widetilde{M}_q^{lr}(W^{q-1})^{-1}\Big]\\
&\qquad\cdot\Big\{Id-\Big[W^{q-1}\widetilde{R}^{1-\cdots-(q-1)}(W^{q-1})^{-1}\Big]\cdot\Big[W^{q-1}\widetilde{M}_q^{ll}(W^{q-1})^{-1}\Big]\Big\}^{-1}\\
&\qquad\cdot\Big[W^{q-1}\widetilde{R}^{1-\cdots-(q-1)}(W^{q-1})^{-1}\Big].
\end{aligned}
\tag{3.19}
$$

Setting

$$A=W^{q-1}\widetilde{M}_q^{ll}(W^{q-1})^{-1},\qquad B=W^q\widetilde{M}_q^{lr}(W^{q-1})^{-1}$$

and combining Lemma 3.6 with the inductive assumption

$$\|\ W^{q-1}\cdot\widetilde{R}^{1-\cdots-(q-1)}\cdot(W^{q-1})^{-1}\ \|_1<1,$$

we get (3.18) by (3.19) and thus complete the proof of (3.17).

We now prove inductively that the BV stability condition (3.1), (3.2) implies (3.8). For $m=2$, the conditions (3.12) and (3.14) are by Lemmas 3.3 and 3.4 equivalent to

$$\text{all eigenvalues of } \widetilde{M}_1^{rr}\cdot\widetilde{M}_2^{ll} \text{ are } <1 \text{ in absolute value.} \tag{3.20}$$

However,

$$\text{Spec } M_1^{rr}M_2^{ll}\subset\text{Spec }\widetilde{M}_1^{rr}\widetilde{M}_2^{ll}\subset\big(\text{Spec }M_1^{rr}M_2^{ll}\big)\cup\{0\},$$

thus (3.20) is equivalent to

$$\text{spRad}(F^{1-2})<1,$$

which is in turn precisely the condition (3.8).

Note that we proved above even more than we need to at this point—we proved the equivalence of (3.1), (3.2), and (3.8) in case $m=2$ of only two large shocks present.

Let now $m>2$. Since (3.13) for $q=m-1$ implies

$$\|\ W^{q-2}\widetilde{M}_{q-1}^{ll}(W^{q-1})^{-1}\ \|_1<1,$$

by the inductive assumption we get

$$\text{spRad}(F^{1-\cdots-q})<1\quad\forall q:2\dots m-1.$$

However, by (3.14) and (3.17) for $q=m-1$, in view of Lemma 3.4 and definition (3.9)

$$\|\ W^{m-1}\widetilde{F}^{1-\cdots-m}(W^{m-1})^{-1}\ \|_1<1,$$

which by Lemma 3.3 implies finally

$$\mathrm{spRad}(F^{1-\cdots-m}) < 1.$$

This finishes the proof of (3.1), (3.2) $\Rightarrow$ (3.8). □

Step 2. (3.8) $\Rightarrow$ (3.1), (3.2). We use the equivalent form of the BV stability condition (3.1), (3.2) given in Lemma 3.5(ii).

We proceed by induction on m. For $m = 2$ the assertion has already been established in Step 1. Let $m > 2$ and fix $\lambda \geq 1$. We will show that

$$\det(Odd_m \cdot Even_m - \lambda Id) \neq 0, \tag{3.21}$$

which by the property of nonnegative matrices mentioned in the proof of Lemma 3.3 will prove the theorem.

Assume first that m is an odd number. By known formulae on the determinant of block matrices (see [G]) and a few easy computations one gets

$$\begin{aligned}
&\det(Odd_m \cdot Even_m - \lambda Id)\\
&\quad = \det(Odd_{m-1} \cdot Even_{m-1} - \lambda Id)\\
&\qquad \cdot \det\Big(\widetilde{M}_m^{ll}\widetilde{M}_{m-1}^{rr} + \widetilde{M}_m^{ll} \cdot A_m \cdot (\lambda Id - Odd_{m-1} \cdot Even_{m-1})^{-1}\\
&\qquad\qquad \cdot B_m \cdot \widetilde{M}_{m-1}^{rl} - \lambda Id\Big),
\end{aligned} \tag{3.22}$$

where A_m is an $n \times ((m-2)\cdot n)$ block matrix of the form

$$A_m = \begin{bmatrix} 0 & \dots & \dots & 0 & \widetilde{M}_{m-1}^{lr} \end{bmatrix},$$

and B_m is an $((m-2)\cdot n) \times n$ block matrix

$$B_m = \begin{bmatrix} 0 & \dots & 0 & \widetilde{M}_{m-2}^{rl} & \widetilde{M}_{m-2}^{rr} \end{bmatrix}^T,$$

while Odd_{m-1} and $Even_{m-1}$ are defined analogously to Odd_m and $Even_m$ as in Lemma 3.5(ii).

Note that the Schochet condition (3.8) implies (by the inductive assumption)

$$\det\left(Odd_{m-1} \cdot Even_{m-1} - \lambda Id\right) \neq 0, \tag{3.23}$$

$$\mathrm{spRad}(F^{1-\cdots-m}) < 1. \tag{3.24}$$

By the definitions (3.9)–(3.11)

$$F^{1-\cdots-m} = M_m^{ll} \cdot \Big[M_{m-1}^{rr} + M_{m-1}^{lr}\big(Id - F^{1-\cdots-(m-1)}\big)^{-1} \cdot R^{1-\cdots-(m-2)} M_{m-1}^{rl}\Big].$$

Thus, in view of (3.23) and (3.24), the needed (3.21) will follow from (3.22) provided that

$$\begin{aligned}
&A_m \cdot (Id - Odd_{m-1} \cdot Even_{m-1})^{-1} \cdot B_m\\
&\qquad = \widetilde{M}_{m-1}^{lr} \cdot \big(Id - \widetilde{F}^{1-\cdots-(m-1)}\big)^{-1} \cdot \widetilde{R}^{1-\cdots-(m-2)}.
\end{aligned} \tag{3.25}$$

By the same kind of reasoning it is possible to prove that for m even, (3.21) is a consequence of the formula

$$\begin{aligned} C_m \cdot (Id - Odd_{m-1} \cdot Even_{m-1})^{-1} \cdot D_m \\ = \left(Id - \widetilde{F}^{1-\cdots-(m-1)}\right)^{-1} \cdot \widetilde{R}^{1-\cdots-(m-2)} \cdot \widetilde{M}^{rl}_{m-1}, \end{aligned} \tag{3.26}$$

where C_m is an $n \times ((m-2)\cdot n)$ block matrix of the form

$$C_m = \left[\begin{array}{ccccc} 0 & \dots & 0 & \widetilde{M}^{lr}_{m-2} & \widetilde{M}^{rr}_{m-2} \end{array} \right],$$

and D_m is an $((m-2)\cdot n) \times n$ block matrix

$$D_m = \left[\begin{array}{ccccc} 0 & \dots & \dots & 0 & \widetilde{M}^{rl}_{m-1} \end{array} \right]^T.$$

In the remaining part of the proof we will concentrate on showing that (3.25) holds for every odd number m. The proof of (3.26) is entirely the same, so we leave it to the careful reader.

We are going to prove (3.25) by induction on odd numbers m. For $m = 3$, the left-hand side of (3.25) reduces to

$$\widetilde{M}^{lr}_2 \cdot (Id - \widetilde{M}^{rr}_1 \cdot \widetilde{M}^{ll}_2)^{-1} \cdot \widetilde{M}^{rr}_1,$$

which is precisely equal to $\widetilde{M}^{lr}_2 \cdot \left(Id - \widetilde{F}^{1-2}\right)^{-1} \cdot \widetilde{R}^1$ by (3.6) and (3.9).

For $m > 3$ and odd, computing $(Id - Odd_{m-1} \cdot Even_{m-1})^{-1}$ in terms of the matrices $Odd_{m-3}, Even_{m-3}$, and the basic block-interaction matrices M^{xy}_q, we receive the equivalent form of the left-hand side of the formula (3.25):

$$\begin{aligned} A_m \cdot (Id - Odd_{m-1} \cdot Even_{m-1})^{-1} \cdot B_m \\ = \left[\begin{array}{cc} 0 & \widetilde{M}^{lr}_{m-1} \end{array} \right] \\ \cdot \left\{ Id - \left[\begin{array}{cc} \widetilde{M}^{ll}_{m-2} & \widetilde{M}^{rl}_{m-2} \\ \widetilde{M}^{lr}_{m-2} & \widetilde{M}^{rr}_{m-2} \end{array} \right] \cdot \left[\begin{array}{cc} \widetilde{M}^{rr}_{m-3} & 0 \\ 0 & \widetilde{M}^{ll}_{m-1} \end{array} \right] \right. \\ - \left[\begin{array}{c} \widetilde{M}^{ll}_{m-2} \\ \widetilde{M}^{lr}_{m-2} \end{array} \right] \cdot A_{m-2} \cdot (Id - Odd_{m-3} \cdot Even_{m-3})^{-1} \\ \left. \cdot B_{m-2} \cdot \left[\begin{array}{cc} \widetilde{M}^{rl}_{m-3} & 0 \end{array} \right] \right\}^{-1} \cdot \left[\begin{array}{c} \widetilde{M}^{rl}_{m-2} \\ \widetilde{M}^{rr}_{m-2} \end{array} \right]. \end{aligned} \tag{3.27}$$

Using the inductive assumption and the definition (3.11) we reformulate the right-hand side of (3.27):

$$
\begin{aligned}
&A_m \cdot (Id - Odd_{m-1} \cdot Even_{m-1})^{-1} \cdot B_m \\
&\quad = \begin{bmatrix} 0 & \widetilde{M}^{lr}_{m-1} \end{bmatrix} \\
&\qquad \cdot \Bigg\{ Id - \begin{bmatrix} \widetilde{M}^{ll}_{m-2} & \widetilde{M}^{rl}_{m-2} \\ \widetilde{M}^{lr}_{m-2} & \widetilde{M}^{rr}_{m-2} \end{bmatrix} \cdot \begin{bmatrix} \widetilde{M}^{rr}_{m-3} & 0 \\ 0 & \widetilde{M}^{ll}_{m-1} \end{bmatrix} \\
&\qquad\qquad - \begin{bmatrix} \widetilde{M}^{ll}_{m-2} \\ \widetilde{M}^{lr}_{m-2} \end{bmatrix} \cdot \widetilde{M}^{lr}_{m-3} \cdot (Id - \widetilde{F}^{1-\cdots-(m-3)})^{-1} \\
&\qquad\qquad\qquad \cdot \widetilde{R}^{1-\cdots-(m-4)} \begin{bmatrix} \widetilde{M}^{rl}_{m-3} & 0 \end{bmatrix} \Bigg\}^{-1} \cdot \begin{bmatrix} \widetilde{M}^{rl}_{m-2} \\ \widetilde{M}^{rr}_{m-2} \end{bmatrix} \\
&\quad = \begin{bmatrix} 0 & \widetilde{M}^{lr}_{m-1} \end{bmatrix} \cdot \Bigg\{ Id - \begin{bmatrix} \widetilde{M}^{ll}_{m-2} & \widetilde{M}^{rl}_{m-2} \\ \widetilde{M}^{lr}_{m-2} & \widetilde{M}^{rr}_{m-2} \end{bmatrix} \\
&\qquad \cdot \begin{bmatrix} \widetilde{M}^{rr}_{m-3} + & \\ \widetilde{M}^{lr}_{m-3}(Id - \widetilde{F}^{1-\cdots-(m-3)})^{-1} \cdot & 0 \\ \cdot \widetilde{R}^{1-\cdots-(m-4)} \widetilde{M}^{rl}_{m-3} & \\ 0 & \widetilde{M}^{ll}_{m-1} \end{bmatrix} \Bigg\}^{-1} \cdot \begin{bmatrix} \widetilde{M}^{rl}_{m-2} \\ \widetilde{M}^{rr}_{m-2} \end{bmatrix} \\
&\quad = \begin{bmatrix} 0 & \widetilde{M}^{lr}_{m-1} \end{bmatrix} \cdot \Bigg\{ Id - \begin{bmatrix} \widetilde{M}^{ll}_{m-2} & \widetilde{M}^{rl}_{m-2} \\ \widetilde{M}^{lr}_{m-2} & \widetilde{M}^{rr}_{m-2} \end{bmatrix} \\
&\qquad\qquad \cdot \begin{bmatrix} \widetilde{R}^{1-\cdots-(m-3)} & 0 \\ 0 & \widetilde{M}^{ll}_{m-1} \end{bmatrix} \Bigg\}^{-1} \cdot \begin{bmatrix} \widetilde{M}^{rl}_{m-2} \\ \widetilde{M}^{rr}_{m-2} \end{bmatrix}.
\end{aligned} \tag{3.28}
$$

Calling

$$
\begin{aligned}
X &= Id - \widetilde{M}^{ll}_{m-2} \widetilde{R}^{1-\cdots-(m-3)}, \\
Y &= -\widetilde{M}^{rl}_{m-2} \widetilde{M}^{ll}_{m-1}, \\
Z &= -\widetilde{M}^{lr}_{m-2} \widetilde{R}^{1-\cdots-(m-3)}, \\
W &= Id - \widetilde{M}^{lr}_{m-2} \widetilde{M}^{ll}_{m-1},
\end{aligned}
$$

we rewrite the right-hand side of (3.28):

$$
\begin{aligned}
&\begin{bmatrix} 0 & \widetilde{M}^{lr}_{m-1} \end{bmatrix} \cdot \begin{bmatrix} X & Y \\ Z & W \end{bmatrix}^{-1} \cdot \begin{bmatrix} \widetilde{M}^{rl}_{m-2} \\ \widetilde{M}^{rr}_{m-2} \end{bmatrix} \\
&\qquad = \widetilde{M}^{lr}_{m-1} \cdot \Big(-(W - ZX^{-1}Y)^{-1} ZX^{-1} \cdot \widetilde{M}^{rl}_{m-2} \\
&\qquad\qquad + (W - ZX^{-1}Y)^{-1} \cdot \widetilde{M}^{rr}_{m-2} \Big) \\
&\qquad = \widetilde{M}^{lr}_{m-1} \cdot (W - ZX^{-1}Y)^{-1} \cdot \big(\widetilde{M}^{rr}_{m-2} - ZX^{-1} \cdot \widetilde{M}^{rl}_{m-2} \big) \\
&\qquad = \widetilde{M}^{lr}_{m-1} \cdot (Id - \widetilde{R}^{1-\cdots-(m-2)} \widetilde{M}^{ll}_{m-1})^{-1} \cdot \widetilde{R}^{1-\cdots-(m-2)},
\end{aligned} \tag{3.29}
$$

because, by definitions (3.9)–(3.11)

$$W - ZX^{-1}Y = Id - \widetilde{R}^{1-\cdots-(m-2)}\widetilde{M}^{ll}_{m-1},$$
$$\widetilde{M}^{rr}_{m-2} - ZX^{-1}\cdot\widetilde{M}^{rl}_{m-2} = \widetilde{R}^{1-\cdots-(m-2)}.$$

The equality (3.29) together with (3.28) prove (3.25). The proof of Step 2 and thus also the proof of Theorem 3.2 is complete. □

3.5. L^1 stability condition. In connection with (3.3) and (3.4), we define the matrices $N^{rr}_q, N^{rl}_q, N^{lr}_q$, and N^{ll}_q $(q : 1\dots m)$, having the same dimensions as their corresponding matrices M^{xy}_q in (3.5), and with their (nonnegative) entries given by

$$b_{sk} = a_{sk}\cdot\left|\frac{\lambda_s(u^q_0)-\Lambda^q}{\lambda_k(u^q_0)-\Lambda^q}\right| \quad \text{in } N^{rr}_q,$$
$$b_{sk} = a_{sk}\cdot\left|\frac{\lambda_s(u^{q-1}_0)-\Lambda^q}{\lambda_k(u^q_0)-\Lambda^q}\right| \quad \text{in } N^{rl}_q,$$
$$b_{sk} = a_{sk}\cdot\left|\frac{\lambda_s(u^q_0)-\Lambda^q}{\lambda_k(u^{q-1}_0)-\Lambda^q}\right| \quad \text{in } N^{lr}_q,$$
$$b_{sk} = a_{sk}\cdot\left|\frac{\lambda_s(u^{q-1}_0)-\Lambda^q}{\lambda_k(u^{q-1}_0)-\Lambda^q}\right| \quad \text{in } N^{ll}_q.$$

Using the analysis of the previous subsection, we can now state the following.

PROPOSITION 3.7. *The L^1 stability condition* (3.3), (3.4) *is equivalent to the condition* (3.8)*, where the matrices $F^{1-\cdots-q}$ are defined as in* (3.9)–(3.11)*, with every matrix M^{xy}_q replaced by the corresponding one N^{xy}_q. In particular, for $m = 2$,* (3.8) *reduces to*

$$\begin{aligned}&\textit{the spectral radius of an } n\times n \textit{ matrix}\\ &|\mathcal{S}-\Lambda^1 Id|\cdot\widetilde{M}^{rr}_1\cdot|\mathcal{S}-\Lambda^1 Id|^{-1}\cdot|\mathcal{S}-\Lambda^2 Id|\cdot\widetilde{M}^{ll}_2\cdot|\mathcal{S}-\Lambda^2 Id|^{-1}\\ &\textit{is smaller than } 1,\end{aligned}\tag{3.30}$$

where

$$|\mathcal{S}-\Lambda Id| = \operatorname{diag}(|\lambda_1(u^1_0)-\Lambda|,\dots,|\lambda_n(u^1_0)-\Lambda|).$$

Remark 3.8. It has recently been brought to our attention that conditions similar to our BV and L^1 stability conditions, though expressed in the language of matrix analysis, can be found in the book [LY].

The authors investigate the (short time) existence and regularity of classical solutions to the so-called typical boundary value problems on fan-shaped domains for quasi-linear hyperbolic systems with smooth coefficients. In particular, they show the existence of a unique C^1 solution to this problem, provided that the so-called minimal characterizing number of the characterizing matrix for the typical boundary value problem is smaller than 1 (Theorem 1.1 in Chapter 4). If the same holds for the second characterizing matrix (see paragraph 4 in Chapter 7), then the corresponding solution is C^2 regular (Theorem 1.1 in Chapter 7).

These results can well be applied to the quasi-linear system (1.4) with the boundary conditions (1.5) along the boundaries of the angular domains given by the large

shocks in the solution of (1.1), (1.2), (1.3). The boundary conditions (1.5) appear already in the solvable form (see Lemma 5.10 in Chapter 2), that is, some of the components of u at the vertex $x = 0, t = 0$ (namely, the components corresponding to the outgoing modes) are explicitly expressed as functions of the others (corresponding to the incoming modes). It is not hard to notice that the characterizing matrix of this problem is made up of the quantities $\{\frac{\partial}{\partial \epsilon_k^{in}}\ \epsilon_s^{out}\}$ in such a way that its minimal characterizing number is smaller than 1 iff our BV stability condition holds. In a similar manner, the mentioned solvability condition for the second characterizing matrix, containing the numbers $\{\frac{\partial}{\partial \epsilon_k^{in}}\left(\frac{\epsilon_s^{out}\cdot(\lambda_s^{out}-\Lambda^q)}{(\lambda_k^{in}-\Lambda^q)}\right)\}$, is equivalent to our L^1 stability condition.

The results in [LY] thus imply the local in time existence of the piecewise C^1 (respectively, C^2) solution to the problem (1.1), (1.2) with $\bar{u}$ smooth except at the point $x = 0$, where it induces the Riemann problem "close" to (u^-, u^+).

4. Systems of two equations. In the particular case $n = m = 2,\ i_1 = 1,\ i_2 = 2$, the matrices M_1^{rr} and M_2^{ll} reduce to single numbers, and the L^1 stability condition (3.30) appears in a simple form:

$$\left|\frac{\partial \epsilon_2^{out}}{\partial \epsilon_1^{in}}\ |\epsilon_1^{in}=0\right| \cdot \left|\frac{\partial \epsilon_1^{out}}{\partial \epsilon_2^{in}}\ |\epsilon_2^{in}=0\right| \cdot \frac{\lambda_1(u_0^1)-\Lambda^2}{\lambda_1(u_0^1)-\Lambda^1}\cdot\frac{\lambda_2(u_0^1)-\Lambda^1}{\lambda_2(u_0^1)-\Lambda^2} < 1. \tag{4.1}$$

Similarily, the BV stability condition (3.1), (3.2) is equivalent to

$$\left|\frac{\partial \epsilon_2^{out}}{\partial \epsilon_1^{in}}\ |\epsilon_1^{in}=0\right| \cdot \left|\frac{\partial \epsilon_1^{out}}{\partial \epsilon_2^{in}}\ |\epsilon_2^{in}=0\right| < 1. \tag{4.2}$$

In both (4.1) and (4.2) the first derivative corresponds to the right interaction with the large shock of the first family, while the second derivative corresponds to the left interaction with the large shock of the second characteristic family.

In what follows we show that (4.1) and (4.2) are equivalent, respectively, to the appropriate conditions providing stability results in [BC] and [W].

4.1. The Bressan–Colombo L^1 stability condition [BC]. In the setting of [BC],

$$\kappa_1 = \frac{\partial \epsilon_2^{out}}{\partial \epsilon_1^{in}}\ |\epsilon_1^{in}=0 = -\frac{\left\langle \frac{\partial \Psi^2(u_0^0,u_0^1)}{\partial u^1}, r_1(u_0^1)\right\rangle}{\left\langle \frac{\partial \Psi^2(u_0^0,u_0^1)}{\partial u^1}, r_2(u_0^1)\right\rangle}$$

and

$$\kappa_2 = \frac{\partial \epsilon_1^{out}}{\partial \epsilon_2^{in}}\ |\epsilon_2^{in}=0 = -\frac{\left\langle \frac{\partial \Psi^1(u_0^1,u_0^2)}{\partial u^1}, r_2(u_0^1)\right\rangle}{\left\langle \frac{\partial \Psi^1(u_0^1,u_0^2)}{\partial u^1}, r_1(u_0^1)\right\rangle},$$

where

$$\Psi^1(u^1,u^2) = \langle l_1(u^1,u^2), u^1-u^2\rangle,$$

$$\Psi^2(u^0,u^1) = \langle l_2(u^0,u^1), u^0-u^1\rangle,$$

l_1 and l_2 being the left eigenvectors of the averaged flux gradient matrix between the reference points u.

One sees that the Bressan–Colombo stability condition

$$\left|\kappa_1 \cdot \frac{\lambda_1(u_0^1) - \Lambda^2}{\lambda_1(u_0^1) - \Lambda^1}\right| \cdot \left|\kappa_2 \cdot \frac{\lambda_2(u_0^1) - \Lambda^1}{\lambda_2(u_0^1) - \Lambda^2}\right| < 1$$

is precisely (4.1).

4.2. The Wang BV stability condition [W]. In [W], (1.1), (1.7) is assumed to satisfy the following finiteness condition:

Let

$$\begin{aligned} \left(\Lambda^1 Id - Df(u_0^1)\right)^{-1}(u_0^1 - u_0^0) &= \alpha r_1(u_0^1) + \beta r_2(u_0^1), \\ \left(Df(u_0^1) - \Lambda^2 Id\right)^{-1}(u_0^2 - u_0^1) &= \gamma r_1(u_0^1) + \delta r_2(u_0^1). \end{aligned} \tag{4.3}$$

Then

$$|\beta\gamma| < |\alpha\delta|. \tag{4.4}$$

The above condition is a reduction of a multidimensional BV stability condition (to be found in [Me]) to the case of one space dimension.

THEOREM 4.1. *Assume that both shocks in the reference solution* (1.7) *(recall that* $m = 2$*) are Majda stable and Lax admissible. Then the condition* (4.4) *is equivalent to the BV stability condition* (4.2).

Proof. It is enough to show that in the context of (4.3), (4.4), (4.2), there hold

$$\left|\frac{\beta}{\alpha}\right| = \left|\frac{\partial \epsilon_2^{out}}{\partial \epsilon_1^{in}} \; |\epsilon_1^{in} = 0\right|, \tag{4.5}$$

$$\left|\frac{\gamma}{\delta}\right| = \left|\frac{\partial \epsilon_1^{out}}{\partial \epsilon_2^{in}} \; |\epsilon_2^{in} = 0\right|. \tag{4.6}$$

We focus on (4.5) and thus the case when the large shock (u_0^0, u_0^1) is hit from the right by a small wave of the first characteristic family and strength ϵ_1^{in}. The proof of (4.6) is entirely similar, so we omit it.

Let $F : \Omega^0 \times \Omega^1 \times I \longrightarrow \mathbf{R}$ be defined as follows (I is here a small neighborhood of $0 \in \mathbf{R}$):

$$F(u^-, u^+, \epsilon) = \Psi_0(u^-, \widetilde{\Phi}_2(u^+, \epsilon)),$$

where Ψ_0 is as in (2.5), (2.6) (its existence is implied by the proof of Theorem 2.1, in view of the Majda stability of the first large shock). The functions $\widetilde{\Phi}_i : \Omega^1 \times I \longrightarrow \Omega^1$ for $i = 1, 2$ are such that

$$\widetilde{\Phi}_i(u^+, \epsilon) = u^- \quad \text{iff} \quad \Phi_i(u^-, \epsilon) = u^+,$$

where $\Phi_i : \Omega^1 \times I \longrightarrow \Omega^1$ for a fixed u^- coincides with the ith rarefaction curve in the positive part of I, and for $\epsilon \in I$ negative follows the ith shock curve through the argument point u (compare [L]). It is not hard to notice that $\frac{\partial}{\partial \epsilon}\widetilde{\Phi}_i(u, 0) = -r_i(u)$.

The fundamental equation relating the strengths ϵ_1^{in} and ϵ_2^{out} in (4.5) has by (2.6) the form

$$F(u_0^0, \Phi_1(u_0^1, \epsilon_1^{in}), \epsilon_2^{out}) = 0. \tag{4.7}$$

Differentiating (4.7) with respect to ϵ_1^{in} at $\epsilon_1^{in} = 0$ and using (2.13), we receive

$$\begin{aligned} 0 &= \frac{\partial \Phi_0}{\partial u^1}(u_0^0, u_0^1) \cdot r_1(u_0^1) - \frac{\partial \Phi_0}{\partial u^1}(u_0^0, u_0^1) \cdot r_2(u_0^1) \cdot \frac{\partial \epsilon_2^{out}}{\partial \epsilon_1^{in}} \ |\epsilon_1^{in} = 0 \\ &= V_1(u_0^1 - u_0^0)^T \cdot \big[Df(u_0^1) - \Lambda^1 Id\big] \cdot \left(r_1(u_0^1) - r_2(u_0^1) \cdot \frac{\partial \epsilon_2^{out}}{\partial \epsilon_1^{in}} \ |\epsilon_1^{in} = 0 \right). \end{aligned} \tag{4.8}$$

Since $V_1(u_0^1 - u_0^0)$ is orthogonal to $u_0^1 - u_0^0$, (4.8) is equivalent to

$$\big[Df(u_0^1) - \Lambda^1 Id\big] \cdot \left(r_1(u_0^1) - r_2(u_0^1) \cdot \frac{\partial \epsilon_2^{out}}{\partial \epsilon_1^{in}} \ |\epsilon_1^{in} = 0 \right) = s \cdot (u_0^1 - u_0^0), \tag{4.9}$$

with some $s \neq 0$, as Λ^1 is not an eigenvalue of $Df(u_0^1)$. The first formula in (4.3) is equivalent to

$$\big[Df(u_0^1) - \Lambda^1 Id\big] \cdot \big(-\alpha r_1(u_0^1) - \beta r_2(u_0^1)\big) = (u_0^1 - u_0^0),$$

and thus by (4.9) we get (4.5). □

Acknowledgment. We wish to thank an anonymous referee for bringing to our attention the book [LY].

REFERENCES

[B] A. Bressan, *Hyperbolic systems of conservation laws*, Rev. Mat. Complut., 12 (1999), pp. 135–200.

[BC] A. Bressan and R.M. Colombo, *Unique solutions of* 2×2 *conservation laws with large data*, Indiana Univ. Math. J., 44 (1995), pp. 677–725.

[BLY] A. Bressan, T.P. Liu, and T. Yang, *L^1 stability estimates for $n \times n$ conservation laws*, Arch. Rational Mech. Anal., 149 (1999), pp. 1–22.

[BM] A. Bressan and A. Marson, *A variational calculus for discontinuous solutions of systems of conservation laws*, Comm. Partial Differential Equations, 20 (1995), pp. 1491–1552.

[D] C. Dafermos, *Hyperbolic Conservation Laws in Continuum Physics*, Springer-Verlag, New York, 1999.

[G] F.R. Gantmacher, *Théorie des matrices (tome 1)*, Collection Univ. de Mathématiques, Dunod, Paris, 1966.

[L] P. Lax, *Hyperbolic systems of conservation laws* II, Comm. Pure Appl. Math., 10 (1957), pp. 537–566.

[Le] M. Lewicka, *L^1 stability of patterns of noninteracting large shock waves*, Indiana Univ. Math. J., to appear.

[LT] M. Lewicka and K. Trivisa, *On the L^1 well posedness of systems of conservation laws near solutions containing two large shocks*, J. Differential Equations, to appear.

[LY] T.-T. Li and W.-C. Yu, *Boundary Value Problems for Quasilinear Hyperbolic Systems*, Duke University Press, 1985.

[M] A. Majda, *The stability of multi-dimensional shock fronts*, Mem. Amer. Math. Soc., 41 (1983), no. 275.

[Me] G. Métivier, *Interaction de deux chocs pour un système de deux lois de conservation, en dimension deux d'espace*, Trans. Amer. Math. Soc., 296 (1986), pp. 431–479.

[Scho] S. Schochet, *Sufficient conditions for local existence via Glimm's scheme for large BV data*, J. Differential Equations, 89 (1991), pp. 317–354.

[Sm] J. Smoller, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, New York, 1994.

[W] Y. Wang, *Highly oscillatory shock waves*, in Nonlinear Theory of Generalized Functions– Proceedings of the Workshop Nonlinear Theory of Nonlinear Functions, Vienna 1997, Chapman & Hall/CRC, Boca Raton, FL, 1999, pp. 153–161.

SIAM J. MATH. ANAL.
Vol. 32, No. 5, pp. 1117–1141

THE HEAT EQUATION IN $L_q((0,T),L_p)$-SPACES WITH WEIGHTS*

N. V. KRYLOV†

Abstract. Existence and uniqueness theorems are presented for the heat equation in L_p-spaces with or without weights allowing derivatives of solutions to blow up near the boundary. It is allowed for the powers of summability with respect to space and time variables to be different.

Key words. Sobolev spaces with weights, parabolic equations

AMS subject classification. 35K20

PII. S0036141000372039

We are going to investigate the equation

$$u_t(t,x) = a^{ij}(t)u_{x^ix^j}(t,x) + f(t,x) \tag{0.1}$$

in several subdomains of $\mathsf{R}\times\mathsf{R}^d = \{(t,x) : t \in \mathsf{R}, x \in \mathsf{R}^d\}$ with a^{ij} being only bounded measurable functions of $t \in \mathsf{R}$ satisfying the uniform ellipticity condition. In (0.1) and throughout the article Einstein's summation convention is enforced. We assume that $a(t) := (a^{ij}(t))$ is a symmetric matrix-valued function depending only on t and, for some constants $K, \delta > 0$ and all $t \in \mathsf{R}$ and $\lambda \in \mathsf{R}^d$, we have

$$K|\lambda|^2 \geq a^{ij}(t)\lambda^i\lambda^j \geq \delta|\lambda|^2. \tag{0.2}$$

Equation (0.1) is understood in the sense of generalized functions only with respect to x. In other words, say, in the case of the whole $\mathsf{R}\times\mathsf{R}^d$ by a solution of (0.1), we mean a function $u(t)$, $t \in \mathsf{R}$, taking values in the set of generalized functions on R^d such that, for any $t, s \in \mathsf{R}$ satisfying $t \geq s$ and test function $\varphi \in C_0^\infty(\mathsf{R}^d)$, we have

$$(u(t),\varphi) = (u(s),\varphi) + \int_s^t [a^{ij}(r)(u(r),\varphi_{x^ix^j}) + (f(r),\varphi)]\,dr.$$

The emphasis is on proving solvability in function spaces of Sobolev type with different powers of summability p and q with respect to x and t. This issue arose from the theory of stochastic partial differential equations in domains in Sobolev spaces with weights and it turns out that, in this theory, the spaces with weights are the only reasonable ones where one can look for solutions to equations in domains.

Surprisingly enough, to the best of our knowledge, the case $q \neq p$ was never addressed before even for the heat equation in $\mathsf{R}\times\mathsf{R}^d$ without weights. We could only find references [1] and [12], where different powers of summability can be related to the Cauchy problem in $\{t > 0\}$ for $f = 0$. It turns out that in $\mathsf{R}\times\mathsf{R}^d$ the result we need can be obtained quite easily on the basis of a Banach space version of the Calderón–Zygmund theorem (see section 1), which allows one to pass from $q = p$ to $q \neq p$.

*Received by the editors May 9, 2000; accepted for publication (in revised form) October 16, 2000; published electronically February 21, 2001. This work was partially supported by NSF grant DMS–9876586.

http://www.siam.org/journals/sima/32-5/37203.html

†School of Mathematics, 127 Vincent Hall, University of Minnesota, Minneapolis, MN 55455 (krylov@math.umn.edu).

For (0.1) in $\mathsf{R}\times\mathsf{R}^d_+$ with $\mathsf{R}^d_+=\{x=(x^1,x'):x^1>0,x'\in\mathsf{R}^{d-1}\}$ we impose the zero boundary condition and look for solutions in weighted Sobolev spaces with weights allowing the spatial derivatives of solutions to blow up near the boundary $x^1=0$. Our results in this setting extend the corresponding results in [5], where $q=p$.

This time we use again the Calderón–Zygmund theorem starting with the results valid for $q=p$. However, in order to check the conditions of this theorem, we need some nontrivial properties of the heat semigroup in weighted spaces, which we prove in section 2. In section 3, this allows us to get the result for $\mathsf{R}\times\mathsf{R}^d_+$, but only if $a^{1j}\equiv 0$ for $j\geq 2$. Usually, if one proves a sufficiently strong result for the heat equation, the same or very close result can be proved for equations with variable *continuous* coefficients. However, in the main application, which we have in mind, to stochastic partial differential equations from filtering theory, the regularity of a^{ij} in time is hard to control. Therefore, we always deal only with measurable coefficients and, in section 4 after some additional work, we prove our main result for equations in $\mathsf{R}\times\mathsf{R}^d_+$ in full generality.

The arguments of section 4 are based on Lemma 1.5, which also allows us to give a different proof of the result in $\mathsf{R}\times\mathsf{R}^d$ by using the Marcinkiewicz interpolation theorem rather than the Calderón–Zygmund theorem.

In section 5 we prove a general theorem used in section 2 saying, roughly speaking, that whatever estimate is true for the heat equation, is also true for parabolic equations with coefficients depending only on time. This theorem is equally applicable to Sobolev and Hölder spaces.

Finally, it is worth noting that we also give results for the initial value problems. To give the reader a flavor of our results in $\mathsf{R}\times\mathsf{R}^d_+$, we state a particular case of Theorem 3.2.

THEOREM 0.1. *Let $p,q\in(1,\infty)$, $-1<\alpha<p-1$, $T\in(0,\infty)$, and assume that we are given a function $f(t,x)$ defined for $t>0,x\in\mathsf{R}^d_+$ and such that*

$$\int_0^T\left(\int_{\mathsf{R}^d_+}(x^1)^{\alpha+p}|f(t,x)|^p\,dx\right)^{q/p}dt<\infty.$$

Then on $[0,T]\times\bar{\mathsf{R}}^d_+$ there is a unique function u satisfying the heat equation

$$u_t=\Delta u+f\quad in\quad (0,T)\times\mathsf{R}^d_+,$$

vanishing for $t=0$ and for $x^1=0$ in a natural sense and such that

$$\int_0^T\left(\int_{\mathsf{R}^d_+}(x^1)^{\alpha}[|u(t,x)/x^1|^p+|u_x(t,x)|^p+|x^1u_{xx}(t,x)|^p]\,dx\right)^{q/p}dt<\infty.$$

Throughout the paper we fix $a^{ij}(t)$ and the constants $K,\delta>0$, assume (0.2), and use the following notation:

$$u_{x^i}=\frac{\partial u}{\partial x^i}=D_iu,\quad u_x=(u_{x^1},\dots,u_{x^d})=Du,$$

$$u_{x^ix^j}=\frac{\partial^2u}{\partial x^i\partial x^j}=D_{ij}u,\quad u_{xx}=(u_{x^ix^j})_{i,j=1}^d=D^2u,\quad u_t=\frac{\partial u}{\partial t}.$$

1. Main results without weights. Here we consider (0.1) in usual Sobolev spaces. The following theorems may be well known to specialists. However, the author could not find their proofs in the literature and since the proofs will be later used with some modification in a more complicated situation, they are presented here.

Theorem 1.1 is the main result. We give it three proofs. One of them is based on the Calderón–Zygmund theorem. It seems impossible to carry over this approach to the case of stochastic partial differential equations. The second proof uses the fact that $||u||_p^{np}$ can be written as the integral of $|u(x_1)\cdot\ldots\cdot u(x_n)|^p$ over R^{nd}. This allows us to reduce estimating the $L_{np}(L_p)$-norm of $u(t,x)$ to estimating the L_p-norm of the function $u(t,x_1)\cdot\ldots\cdot u(t,x_n)$. It turns out that this device works equally well for stochastic partial differential equations. Finally, the third proof, given in section 3, uses our result for spaces with weights and is designed to show that in a sense the case with weights is more general than the one without weights.

Define

$$Lu = a^{ij}u_{x^ix^j} - u_t, \tag{1.1}$$

$$L_p = L_p(\mathsf{R}^d), \quad H_p^\gamma = (1-\Delta)^{-\gamma/2}L_p, \quad \mathbb{H}_p^{\gamma,q} = L_q(\mathsf{R}, H_p^\gamma),$$

$$\mathbb{H}_p^{\gamma,q}(T) = L_q((0,T), H_p^\gamma), \quad \mathbb{L}_p^q = \mathbb{H}_p^{0,q}, \quad \mathbb{L}_p^q(T) = \mathbb{H}_p^{0,q}(T).$$

It is important to emphasize that for simplicity of notation we also use the same symbols L_p and H_p^γ for spaces of real-valued, vector-valued, and matrix-valued functions defined on R^d. It is well known that, for any $s\in\mathsf{R}$ and $f\in C_0^\infty(\mathsf{R}^d)$, there exists a unique bounded continuous function $u(t,x)$ on $[s,\infty)\times\mathsf{R}^d$ satisfying $Lu(t)=0$ for $t>s$ with initial condition $u(s)=f$. We denote $u(t)=T_{s,t}f$ and recall that, for each s and t, the operator $T_{s,t}$ is written as the convolution of f with a Gaussian density. In particular, $T_{s,t}f$ is infinitely differentiable in x. Finally, for $f\in C_0^\infty(\mathsf{R}\times\mathsf{R}^d)$, let

$$Rf(t) := \int_{-\infty}^t T_{s,t}f(s)\,ds, \quad Af := D^2Rf.$$

Also remember that, for $f\in C_0^\infty(\mathsf{R}\times\mathsf{R}^d)$, the function Rf satisfies $LRf=-f$.

THEOREM 1.1. *Let $q,p\in(1,\infty)$, $\gamma\in\mathsf{R}$. Then the operator A is uniquely extendable to a bounded operator acting in $\mathbb{H}_p^{\gamma,q}$. If we keep the same notation for the extension, then*

$$||D^2Rf||_{\mathbb{H}_p^{\gamma,q}} \le N(\delta,d,q,p)||f||_{\mathbb{H}_p^{\gamma,q}}. \tag{1.2}$$

Let us make precise that in this theorem the statement that A is an operator acting in $\mathbb{H}_p^{\gamma,q}$ means that it maps the space of real-valued functions of class $\mathbb{H}_p^{\gamma,q}$ into the space of matrix-valued functions of class $\mathbb{H}_p^{\gamma,q}$. We allow ourselves such an abuse of language on some occasions in the future as well. Another comment is that the constant N in (1.2) is independent of K (see (0.2)).

Before proving Theorem 1.1 we derive from it the following theorem.

THEOREM 1.2. *Let $q,p\in(1,\infty)$, $T\in(0,\infty)$, and $\gamma\in\mathsf{R}$. Take $\varepsilon>0$, $f\in\mathbb{H}_p^{\gamma,q}(T)$, and $u_0\in H_p^{\gamma+2-2/q+\varepsilon}$. Then in $\mathbb{H}_p^{\gamma+2,q}(T)$ there is a unique solution of (0.1) with the initial condition $u(0)=u_0$. For this solution*

$$||u_{xx}||_{\mathbb{H}_p^{\gamma,q}(T)} \le N(||f||_{\mathbb{H}_p^{\gamma,q}(T)} + ||u_0||_{H_p^{\gamma+2-2/q+\varepsilon}}), \tag{1.3}$$

where $N = N(\delta, d, q, p, T, \varepsilon)$, and if $u_0 = 0$, then N is independent of T. Finally, if $q = p$, one can take $\varepsilon = 0$.

Proof. Since we can apply the operator $(1-\Delta)^{\gamma/2}$ to both sides of (0.1), we need to prove the theorem only for $\gamma = 0$. After that the independence of N of T if $u_0 = 0$ is derived by using self similarity.

Now, we reduce the general situation to the one in which $u_0 = 0$. To do this we have to show only that there is a continuation of u_0 to a function $\bar{u}$ such that the norms of $\bar{u}_{xx}$ and $\bar{u}_t$ in $\mathbb{L}_p^q(T)$ are controlled by $||u_0||_{H_p^{2-2/q+\varepsilon}}$.

Let $T_t, t \geq 0$, be the semigroup associated with Δ in R^d and let $\bar{u}(t) = T_t u_0$. The semigroup $e^{-t}T_t$ has generator $\Delta - 1$, which also generates the scale of spaces H_p^γ. It follows by Theorem 14.11 of [2] that, for $\theta \in [0,1]$, $t > 0$, and $f \in L_p$, we have

$$||T_t f||_{L_p} \leq N e^t t^{-\theta} ||f||_{H_p^{-2\theta}},$$

where $N = N(d, p, \theta)$. By replacing f with $(1-\Delta)u_0$, taking $\theta = 1/q - \varepsilon/2$, and assuming without losing generality that $\varepsilon < 2/q$, we conclude

$$||\bar{u}(t)||_{H_p^2} \leq N e^t t^{-1/q+\varepsilon/2} ||u_0||_{H_p^{2-2/q+\varepsilon}},$$

$$\int_0^T ||\bar{u}(t)||_{H_p^2}^q \, dt \leq N ||u_0||_{H_p^{2-2/q+\varepsilon}}^q.$$

Since $\bar{u}_t = \Delta \bar{u}$, we see that the function $\bar{u}(t)$ possesses the desired properties. If $q = p$, the same holds for $\varepsilon = 0$ as is shown in section 4.3 of [8]. Therefore, in the rest of the proof we take $u_0 = 0$.

Take $f \in C_0^\infty(\mathsf{R}_+ \times \mathsf{R}^d)$ and let $u = Rf$. As we have mentioned before, u is a classical solution of $Lu = -f$ for $t > 0$ and obviously $u(0) = 0$. Owing to Theorem 1.1 and the fact that $u(t)$ for $t \in [0,T]$ is independent of the values of $f(s)$ for $s \geq T$, we get (1.3) and from the equation $Lu = -f$ that $||u_t||_{\mathbb{L}_p^q(T)} \leq N||f||_{\mathbb{L}_p^q(T)}$. In particular, for any $T < \infty$,

$$\sup_{t \leq T} ||u(t,\cdot)||_{L_p} \leq N||f||_{\mathbb{L}_p^q(T)} \tag{1.4}$$

with N independent of f. It follows that, if f is in a bounded set in $\mathbb{L}_p^q(T)$, then u is in a bounded set in $\mathbb{H}_p^{2,q}(T) \cap C([0,T], L_p)$. Using obvious approximations proves our assertion about the existence of solutions.

To prove uniqueness, let $u \in \mathbb{H}_p^{2,q}(T)$ satisfy $Lu = 0$ in $(0,T) \times \mathsf{R}^d$ and $u(0) = 0$. (Remember that $u(t)$ is weakly continuous in t by the definition of solution.) Then, of course, $u_t = a^{ij} u_{x^i x^j} \in \mathbb{L}_p^q(T)$ and one can find a sequence of infinitely differentiable functions $u_n(t,x)$ vanishing for large $|x|$ and for $t = 0$ such that

$$||u - u_n||_{\mathbb{H}_p^{2,q}(T)} + ||u - u_{nt}||_{\mathbb{L}_p^q(T)} \to 0.$$

In that case, $Lu_n \to Lu = 0$ in $\mathbb{L}_p^q(T)$. Since $u_n = -RLu_n$, (1.4) implies that

$$\sup_{t \leq T} ||u_n(t,\cdot)||_{L_p} \to 0, \quad \sup_{t \leq T} ||u(t,\cdot)||_{L_p} = 0.$$

The theorem is proved. □

The following corollary of Theorem 1.2 is obtained by odd continuation of the functions involved.

COROLLARY 1.3. *All assertions of Theorem* 1.2 *hold true for* $\gamma = 0$ *if we replace* R^d *with* R^d_+ *everywhere, assume that* $a^{1j} \equiv 0$ *for* $j = 2, \dots, d$, *and supplement* (0.1) *with zero boundary condition at* $x^1 = 0$.

To prove Theorem 1.1, we use the following Banach space version of the Calderón–Zygmund theorem. This is a standard result which is discussed, for instance, in Chapter 1 of [10] and can be extracted from more general results of [1]. For a Hilbert space version of this theorem in the form of multipliers along with a version of Theorem 1.1 for $q = p$ and different operators L, we refer the reader to [9].

THEOREM 1.4. *Let* F *and* G *be Banach spaces, let* $p \in (1, \infty)$, *and let* $A : L_p(\mathsf{R}^n, F) \to L_p(\mathsf{R}^n, G)$ *be a linear bounded operator. Assume that if a bounded strongly measurable* F*-valued function* f *has compact support* Γ, *then, for almost any* $x \notin \Gamma$, *we have*

$$Af(x) = \int_{\mathsf{R}^n} K(x,y) f(y)\, dy,$$

where $K(x,y)$ *is a bounded operator from* F *into* G, *defined for* $x \neq y$, *strongly measurable with respect to* y *with norm bounded in* y *outside any neighborhood of* x. *Also assume that* $K(x,y)$ *is strongly measurable with respect to* x *and there exists a constant* N *such that*

$$\int_{|x-y|>2|y-z|} |K(x,y) - K(x,z)|\, dx \leq N$$

for any y *and* z, *which holds, for instance, if* $K(x,y)$ *is weakly differentiable in* y *and* $|\nabla_y K(x,y)| \leq N|x-y|^{-d-1}$.

Then the operator A *is uniquely extendable to a bounded operator from* $L_q(\mathsf{R}^n, F)$ *to* $L_q(\mathsf{R}^n, G)$ *for any* $q \in (1, p]$ *and* A *is of weak-type* $(1,1)$ *on bounded functions with compact support.*

The first proof of Theorem 1.1. In [3] a general theorem is proved which, roughly speaking, says that whatever estimate is true for the heat equation in translation invariant spaces is also true with the same constant for (0.1) with the coefficients depending only on t provided $(a^{ij}(t)) \geq (\delta^{ij})$. (We prove a similar result for equations in half spaces in Theorem 5.1 and show how to use it in the proof of Theorem 2.5.) Of course, we can achieve the inequality $(a^{ij}(t)) \geq (\delta^{ij})$ by using dilations once we are given (0.2). Therefore, we may and will assume that $a^{ij} \equiv \delta^{ij}$. Also as in the proof of Theorem 1.2, assuming $\gamma = 0$ does not restrict generality.

Now we are ready to use Theorem 1.4. From section 4.3 of [8] we know that A is uniquely defined and is bounded as an operator acting in $L_p(\mathsf{R}, L_p)$. We are going to check that A satisfies the assumptions of Theorem 1.4 with $F = G = L_p$.

Observe the simple fact that, for $t > 0$, $k = 1, 2, \dots$, and $f \in L_p$, we have $\partial T_t f/\partial t = \Delta T_t f$ and $||\partial^k T_t f/\partial t^k||_{L_p} \leq N t^{-k} ||f||_{L_p}$, where N depends only on d and k. For $t > 0$ introduce the operator $K(t) = \Delta T_t$ acting from L_p into L_p with norm bounded by Nt^{-1}, where N is independent of t. For $t \leq 0$, let $K(t) = 0$.

Since $a^{ij} = \delta^{ij}$, we have

$$Rf(t,x) = \int_{-\infty}^{t} T_{t-s} f(s,\cdot)(x)\, ds.$$

In addition, if t is at a distance from the support of f, then differentiating the above formula presents no difficulties and we find

$$Af(t,x) = \int_{-\infty}^{t} \Delta T_{t-s} f(s,\cdot)\, dx = \int_{\mathsf{R}} K(t-s) f(s)(x)\, ds.$$

In order to prove that the assumptions of Theorem 1.4 are satisfied, it remains only to use

$$||\partial K(t-s)f/\partial s||_{L_p} = ||\partial^2 T_{t-s}f/\partial s^2||_{L_p} I_{t>s} \leq N|t-s|^{-2}||f||_{L_p}.$$

By Theorem 1.4, A is well defined and bounded as an operator from $L_q(\mathsf{R}, L_p)$ into itself for $1 < q \leq p$. By considering the adjoint to A, we conclude that A is bounded in $L_q(\mathsf{R}, L_p)$ for any $q, p \in (1, \infty)$. The theorem is proved.

To give a different proof of Theorem 1.1 we prepare two auxiliary results. The first one is an equivalent restatement of the same basic a priori estimate used in the above proof of Theorem 1.1.

LEMMA 1.5. *Let $T \leq \infty$, $p \in (1, \infty)$, and let $u \in L_p((0,T) \times \mathsf{R}^d) = \mathbb{L}_p^p(T)$ be a solution of the equation $Lu = f^{ij}_{x^i x^j}$ with zero initial data and with $f^{ij} \in L_p((0,T) \times \mathsf{R}^d)$. Then*

$$||u||_{\mathbb{L}_p^p(T)} \leq N(d, \delta, p) \sum_{ij} ||f^{ij}||_{\mathbb{L}_p^p(T)}.$$

This lemma follows, for instance, from the results of section 4.3 in [8] up to the fact that there the results are stated for the heat equation or from Theorem 5.1 of [4] up to the assertion that N is independent of T. The latter is obtained in a standard way by using self similarity. In almost the same form as stated, this lemma is proved in the appendix of [11].

In the next lemma we do the first step toward considering the power of summability in t equal to multiples of p. This lemma is also crucial for our investigation of equations in R^d_+ in section 4.

LEMMA 1.6. *Let $T \leq \infty$, $p \in (1, \infty)$, and $n = 1, 2, \ldots$. For $k = 1, \ldots, n$, let $\lambda_k \in (0, \infty)$, $\gamma_k \in \mathsf{R}$, and $u^k \in \mathbb{H}_p^{\gamma_k+2,p}(T)$ be solutions of the equation*

$$u^k_t = a^{ij} u^k_{x^i x^j} + f^k$$

with zero initial data and with $f^k \in \mathbb{H}_p^{\gamma_k,p}(T)$. Denote $\Lambda_k = (\lambda_k - \Delta)^{\gamma_k/2}$. Then

$$\int_0^T \prod_{k=1}^n ||\Lambda_k \Delta u^k(t)||^p_{L_p} \, dt \tag{1.5}$$

$$\leq N \sum_{k=1}^n \int_0^T ||\Lambda_k f^k(t)||^p_{L_p} \prod_{j \neq k} ||\Lambda_j \Delta u^j(t)||^p_{L_p} \, dt,$$

where $N = N(n, d, p, \delta)$.

Proof. By considering $\Lambda_k u^k$ instead of u^k, we see that without loss of generality we may assume $\gamma_k = 0$. In this case define $v^k = \Delta u^k$. For $X = (x_1, \ldots, x_n) \in \mathsf{R}^{nd}$ with $x_i \in \mathsf{R}^d$, define

$$V(t, X) = v^1(t, x_1) \cdot \ldots \cdot v^n(t, x_n).$$

Observe that

$$V_t(t, X) = \mathbb{M} V(t, X) + F(t, X),$$

where $\mathbb{M}V = a^{rs}(V_{x_1^r x_1^s} + \cdots + V_{x_n^r x_n^s})$,

$$F(t,X) = \Delta_{x_i} G^i(t,X), \quad G^i(t,X) = f^i(t,x_i)\prod_{j\neq i} v^j(t,x_j).$$

Hence, by Lemma 1.5

$$||V||_{L_p((0,T)\times \mathsf{R}^{nd})} \leq N\sum_i ||G^i||_{L_p((0,T)\times \mathsf{R}^{nd})},$$

and this is exactly (1.5). The lemma is proved. □

The second proof of Theorem 1.1. As in the first proof, we have only to consider the case $a^{ij} \equiv \delta^{ij}$ and $\gamma = 0$. Also, obviously it suffices to prove (1.2) for $f \in C_0^\infty(\mathsf{R}\times\mathsf{R}^d)$. Without loss of generality we assume that $f(t) = 0$ for $t \leq 0$.

Let $u = Rf$. Then u is a classical solution of $u_t = \Delta u + f$ for $t > 0$ with zero initial condition and, even more than that, $u(s) = 0$ for $s \leq 0$. In addition, it is easy to check that $u \in L_p((0,T), H_p^2)$ for each $T < \infty$.

Next we take $q = np$, where $n = 1, 2, \ldots$. By Lemma 1.6 applied to $u^k = u$ we have

$$||u_{xx}||^{np}_{L_{np}((0,T),L_p)} \leq N\int_0^T ||f(t)||^p_{L_p} ||u_{xx}(t)||^{(n-1)p}_{L_p}\, dt,$$

which by Hölder's inequality yields $||u_{xx}||_{L_{np}((0,T),L_p)} \leq N||f||_{L_{np}((0,T),L_p)}$. By letting $T \to \infty$ we obtain (1.2).

To treat general $q \geq p$, it suffices to use the Marcinkiewicz interpolation theorem. As in the first proof, the case $q \leq p$ is considered by duality. The theorem is proved.

2. Some smoothing properties of the heat semigroup in spaces $H^\gamma_{p,\theta}$. In this section we investigate smoothing properties of solutions to (0.1) in $\mathsf{R}_+ \times \mathsf{R}^d_+$ for $f \equiv 0$. Throughout this section we assume

$$a^{1j}(t) \equiv 0, \quad j = 2, \ldots, d. \tag{2.1}$$

The smoothing is measured in terms of $H^\gamma_{p,\theta}$ spaces introduced in [5] and [6] and recalled briefly below. We fix a function $\zeta \in C_0^\infty(\mathsf{R}_+)$ such that

$$\sum_{n=-\infty}^{\infty} \zeta(e^{n-x}) \geq 1,$$

and for $\theta, \gamma \in \mathsf{R}$, and $p \in (1,\infty)$ we define $H^\gamma_{p,\theta}$ as the space of all distributions u on R^d_+ with finite norm given by

$$||u||^p_{H^\gamma_{p,\theta}} = \sum_{n=-\infty}^{\infty} e^{n\theta} ||u(e^n\cdot)\zeta||^p_{H^\gamma_p}.$$

We write $L_{p,\theta} = H^0_{p,\theta}$. By M^β we denote the operator of multiplying by $(x^1)^\beta$, $M = M^1$. It turns out that for integral γ

$$H^\gamma_{p,\theta} = \{u : M^n D^n u \in L_p(\mathsf{R}^d_+, (x^1)^{\theta-d}\, dx) \quad n = 0, 1, \ldots, \gamma\},$$

where $D^n u$ is the collection of all nth derivatives of u. For other γ one can have an idea about the spaces $H^\gamma_{p,\theta}$ by observing that they are complex interpolation spaces

and the dual to $H^\gamma_{p,\theta}$ is $H^{\gamma'}_{p',\theta'}$ with $\gamma' = -\gamma$, $1/p + 1/p' = 1$, and $\theta/p + \theta'/p' = d$. From other properties of $H^\gamma_{p,\theta}$ which are most often used in this article we point out that the operators MD and DM are bounded operators from $H^\gamma_{p,\theta}$ to $H^{\gamma-1}_{p,\theta}$ and, if $M^{-1}u \in H^\gamma_{p,\theta}$ and $\theta \neq d-1, d-1+p$, then

$$||MD^2u||_{H^{\gamma-2}_{p,\theta}} \leq N||M^{-1}u||_{H^\gamma_{p,\theta}} \leq N||MD^2u||_{H^{\gamma-2}_{p,\theta}}. \tag{2.2}$$

We start with the following general lemma from [7]. We give its proof here for the sake of completeness.

LEMMA 2.1. *Let $g(x,y)$ be a function on $\mathsf{R}^d_+ \times \mathsf{R}^d_+$ satisfying*

$$|g(x,y)| \leq (1+x^1)^\alpha (1+y^1)^\beta e^{-\varepsilon|x-y|^2},$$

where $\varepsilon > 0$, $\alpha, \beta \in \mathsf{R}$ are some constants. Let $1 < p \leq q < \infty$, $\theta \in \mathsf{R}$, and $\delta := (\theta - d)q/p - q(\alpha+\beta)$. Assume

$$d - p/q + p(\alpha+\beta) < \theta < d - 1 + p$$

(so that automatically $\delta > -1$ and $\alpha + \beta < 1 - 1/p + 1/q$). Then, for any $u \in L_{p,\theta}$,

$$\int_{\mathsf{R}^d_+} (x^1)^\delta \left| \int_{\mathsf{R}^d_+} g(x,y)u(y)\,dy \right|^q dx \leq N||u||^q_{L_{p,\theta}}, \tag{2.3}$$

where N is independent of u.

Proof. By using Hölder's inequality we see that

$$\left| \int_{\mathsf{R}^d_+} g(x,y)u(y)\,dy \right| \leq (1+x^1)^\alpha I_1^{1/p}(x) I_2^{(p-1)/p}(x),$$

where

$$\begin{aligned} I_1(x) &= \int_{\mathsf{R}^d_+} |y^1|^{\theta-d}|u(y)|^p e^{-\varepsilon|x-y|^2}\,dy \\ &\leq \int_{\mathsf{R}^{d-1}} \left(\int_0^\infty |y^1|^{\theta-d}|u(y)|^p\,dy^1 \right) e^{-\varepsilon|x'-y'|^2}\,dy' = J(x'), \\ I_2(x) &= I_2(x^1) = \int_{\mathsf{R}^d_+} |y^1|^{(d-\theta)/(p-1)} (1+y^1)^{\beta p/(p-1)} e^{-\varepsilon|x-y|^2}\,dy \\ &= N\int_0^\infty |y^1|^{(d-\theta)/(p-1)} (1+y^1)^{\beta p/(p-1)} e^{-\varepsilon|x^1-y^1|^2}\,dy^1, \end{aligned}$$

where the last integral converges due to $(d-\theta)/(p-1) > -1$. Also, as it is easy to see, this integral behaves like $(x^1)^{(d-\theta+\beta p)/(p-1)}$ when $x^1 \to \infty$. Therefore,

$$\begin{aligned} (1+x^1)^\alpha I_2^{(p-1)/p}(x) &\leq N(1+x^1)^{\alpha+\beta+(d-\theta)/p}, \\ (x^1)^\delta \left| \int_{\mathsf{R}^d_+} g(x,y)u(y)\,dy \right|^q &\leq N(x^1 \wedge 1)^\delta I_1^{q/p}(x). \end{aligned}$$

Thus the left-hand side of (2.3) is less than

$$\int_{\mathsf{R}^{d-1}} J^{q/p}(x')\,dx' + \int_{\mathsf{R}^d} I_1^{q/p}(x)\,dx.$$

To estimate these integrals we notice that I_1 and J are convolutions and the $L_{q/p}$-norm of a convolution is less than the L_1-norm of one function times the $L_{q/p}$-norm of the other. Then we obviously come to (2.3). The lemma is proved. □

LEMMA 2.2. *Let $m \in \{0,1,2,\dots\}$ and $D^m = D_{x^1}^{2l} D_{x'}^{m-2l}$ be an mth derivative operator, where $l \in \{0,1,2,\dots\}$, $D_{x^1}^{2l}$ is the operator of taking $2l$ derivatives in x^1, and $D_{x'}^{m-2l}$ is an $(m-2l)$th derivative with respect to x'. Define*

$$p(t,x) = \frac{1}{(4\pi t)^{d/2}} e^{-|x|^2/(4t)}, \quad Ax = (-x^1, x^2, \dots, x^d),$$

$$p(t,x,y) = p(t,x-y) - p(t,x-Ay), \quad p^{(m)}(t,x,y) = D^m p(t,x,y).$$

Then

$$p^{(m)}(t,x,y) = 0 \quad \text{for} \quad x^1 y^1 = 0, t > 0, \tag{2.4}$$

$$p^{(m)}(t,x,y) = (-1)^m p^{(m)}(t,y,x) \quad \text{for} \quad x, y \in \mathsf{R}_+^d, t > 0. \tag{2.5}$$

Proof. If $l = 0$, (2.4) follows from the equality $p(t,x,y) = 0$ for $x^1 y^1 = 0$. If $l \geq 1$, it suffices to use the induction on l after noticing that, for $x^1 y^1 = 0$, we have

$$\begin{aligned} D_{x^1}^{2l} D_{x'}^{m-2l} p(t,x,y) &= D_{x^1}^{2(l-1)} D_{x'}^{m-2l} \Delta_x p(t,x,y) \\ &= D_t D_{x^1}^{2(l-1)} D_{x'}^{m-2l} p(t,x,y). \end{aligned}$$

To prove (2.5), notice that both sides satisfy the heat equation in x with zero boundary condition on $x^1 = 0$ due to (2.4). Their initial values also coincide since $D_x^m \delta(x-y) = (-1)^m D_x^m \delta(y-x)$. The lemma is proved. □

Denote by $\tilde{T}_t$ the semigroup associated with the operator Δ in R_+^d with zero boundary condition on $\{x^1 = 0\}$. In the following two lemmas we present some properties of the operator

$$S_t^{(m)} = M^{-1} D^m \tilde{T}_t M^{-1},$$

where D^m is taken from Lemma 2.2.

LEMMA 2.3. *Let $1 < p \leq q < \infty$, $\tau, \theta \in \mathsf{R}$, $\sigma := d + (\theta - d)q/p + q\tau$. Assume*

$$d - p/q - p\tau < \theta < d - 1 + p, \quad \tau \leq 2, \quad n \in \{0,1,2,\dots\}.$$

Then, for any $u \in H_{p,\theta}^n$,

$$||S_1^{(m)} u||_{H_{q,\sigma}^n} \leq N ||u||_{H_{p,\theta}^n}, \tag{2.6}$$

where N is independent of u.

Proof. It is well known that

$$\tilde{T}_t u(x) = \int_{\mathsf{R}_+^d} p(t,x,y) u(y)\, dy.$$

Therefore, for $g(x,y) := (x^1 y^1)^{-1} D_x^m p(1,x,y)$, we have

$$S_1^{(m)} u(x) = \int_{\mathsf{R}_+^d} g(x,y) u(y)\, dy.$$

We split the last integral into two parts. Let $\zeta(x^1)$ be an infinitely differentiable function on R such that $\zeta(x^1)=0$ if $0\leq x^1\leq 1$ and $\zeta(x^1)=1$ if $x^1\geq 2$. Denote $\eta=1-\zeta$. Then, for $g_1(x,y):=g(x,y)\zeta(y)$ and $g_2(x,y):=g(x,y)\eta(y)$, we have

$$S_1^{(m)}u(x)=\int_{\mathsf{R}^d_+}g_1(x,y)u(y)\,dy+\int_{\mathsf{R}^d_+}g_2(x,y)u(y)\,dy$$

$$=:S_{11}^{(m)}u(x)+S_{12}^{(m)}u(x)$$

and we estimate each term separately by using Lemma 2.1. Observe that almost obviously, for any derivative D^k with respect to (x,y), we have

$$|D^kp(1,x,y)|\leq Ne^{-|x-y|^2/8},\quad x,y\in\mathsf{R}^d_+. \tag{2.7}$$

Hence, for $x^1\geq 1$, by noticing that the only values of y^1 for which g_1 does not vanish satisfy $y^1\geq 1$, we get

$$|g_1(x,y)|\leq N(1+x^1)^\alpha(1+y^1)^\beta e^{-|x-y|^2/8} \tag{2.8}$$

if $\alpha\geq -1$ and $\beta\geq -1$. Furthermore, (2.8) holds for $x^1\leq 1$ (and all y) as well, since owing to (2.4) we have

$$|g_1(x,y)|=\left|\zeta(y)(y^1)^{-1}\int_0^1 D_{x^1}D_x^mp(1,rx^1,x',y)\,dr\right|$$

and, in addition, $|rx^1-y^1|\geq|x^1-y^1|$ if $x^1\leq 1\leq y^1$. By letting $\tau=-\alpha-\beta$, from Lemma 2.1 we conclude that (2.6) holds with $n=0$ and $S_{11}^{(m)}$ in place of $S_1^{(m)}$.

As long as $S_{12}^{(m)}$ is concerned, notice that

$$\begin{aligned}g_2(x,y)&=\eta(y)(x^1)^{-1}\int_0^1 D_{y^1}D_x^mp(1,x,ry^1,y')\,dr\\&=\eta(y)\int_0^1\int_0^1 D_{y^1}D_{x^1}D_x^mp(1,sx^1,x',ry^1,y')\,drds.\end{aligned} \tag{2.9}$$

The first equality and (2.7) show that, if $x^1\geq 2$ and $y^1\leq 2$,

$$|g_2(x,y)|\leq N(1+x^1)^\alpha(1+y^1)^\beta e^{-|x-y|^2/8} \tag{2.10}$$

with $\alpha\geq -1$ and any β. Inequality (2.10) also holds if $x^1\geq 2$ and $y^1\geq 2$ since in this case $g_2(x,y)=0$.

The second equality in (2.9) shows that (2.10) holds if $0<x^1,y^1<2$. Again trivially it also holds if $x^1\leq 2$ and $y^1\geq 2$. Thus, (2.10) is true for all $x,y\in\mathsf{R}^d_+$ with no restriction on $\tau=-\alpha-\beta$, and Lemma 2.1 allows us to conclude that (2.6) holds with $n=0$ and $S_{12}^{(m)}$ in place of $S_1^{(m)}$. This proves (2.6) for $n=0$.

Now we pass to the case of general $n\geq 1$. Observe that in order to prove that a function f belongs to $H^n_{p,\theta}$, we have to prove that $M^kD^kf\in L_{p,\theta}$ for all $k=0,1,\dots,n$. By using the fact that, for $k\geq 1$,

$$M^kD^kf=M^{k-1}D^kMf+h, \tag{2.11}$$

where h is a linear combination of M^rD^rf with $r \le k-1$, and bearing in mind the induction on n we see that it suffices to consider only terms $M^{k-1}D^kMf$.

Thus, we need only to prove that

$$||M^{n-1}D^nMS_1^{(m)}u||_{L_{q,\sigma}} \le N||u||_{H^n_{p,\theta}}. \tag{2.12}$$

We take the functions ζ and η from above and for any function f on R^d_+ denote $\bar{f}$ its even with respect to x^1 extension to R^d.

Then

$$\begin{aligned} M^{n-1}D^nMS_1^{(m)}u =& M^{n-1}D^mT_1D^n(M^{-1}\bar{\zeta}\bar{u}) \\ &+ M^{n-1}D^{n+m}\tilde{T}_1M^{-1}\eta u = Iu + Ju. \end{aligned} \tag{2.13}$$

First, we estimate J. The kernel of J is

$$\tilde{g}_2(x,y) := \eta(y^1)(x^1)^{n-1}(y^1)^{-1}D_x^{n+m}p(1,x,y).$$

Here, owing to $p(t,x,y) = 0$ for $y^1 = 0$, we have $D_x^{n+m}p(1,x,y) = 0$ if $y^1 = 0$. Upon remembering that $n \ge 1$, we easily derive that $\tilde{g}_2$ satisfies (2.10) with $\alpha \ge n-1$ and any β. Hence, for $\tau = -\alpha - \beta$, by Lemma 2.1 we obtain

$$||Ju||_{L_{q,\sigma}} \le N||u||_{L_{p,\theta}} \le N||u||_{H^n_{p,\theta}}. \tag{2.14}$$

Next, we notice that by Corollary 2.4 of [5], for $v = \zeta u$, we have $||v||_{H^n_{p,\theta}} \le N||u||_{H^n_{p,\theta}}$. Also by Leibnitz's rule, $D^nM^{-1}v$ is written as a linear combination of $M^{-1-k}D^{n-k}v = M^{-1-n}M^{n-k}D^{n-k}v$. It follows that

$$D^n(M^{-1}\zeta u) = M^{-1-n}h, \quad ||h||_{L_{p,\theta}} \le N||u||_{H^n_{p,\theta}},$$

and, in addition, $h(y) = 0$ for $0 \le y^1 \le 1$. Furthermore,

$$\begin{aligned} |Iu(x)| &\le N(x^1)^{n-1}\int_{\mathsf{R}^d_+}(e^{-|x-y|^2/8} + e^{-|x-Ay|^2/8})I_{y^1\ge 1}(y^1)^{-1-n}|h(y)|\,dy \\ &\le N(x^1)^{n-1}\int_{\mathsf{R}^d_+} e^{-|x-y|^2/8}(1+y^1)^{-1-n}|h(y)|\,dy. \end{aligned}$$

We observe that

$$(x^1)^{n-1}(1+y^1)^{-1-n}e^{-|x-y|^2/8} \le N(1+x^1)^{\alpha}(1+y^1)^{\beta}e^{-|x-y|^2/8}$$

if $\alpha = n-1$ and $\beta \ge -n-1$, and by Lemma 2.1 after denoting $\tau = -\alpha - \beta$ we conclude

$$||Iu||_{L_{q,\sigma}} \le N||h||_{L_{p,\theta}} \le N||u||_{H^n_{p,\theta}}.$$

By combining this with (2.13) and (2.14), we finally arrive at (2.12). The lemma is proved. □

LEMMA 2.4. *Let $1 < p \le q < \infty$, $\tau, \theta \in \mathsf{R}$, $\sigma := d + (\theta - d)q/p + q\tau$, and*

$$d - p/q - p\tau < \theta < d - 1 + p.$$

(i) *If* $\tau \leq 1$ *and* $n \in \{1,2,3,\ldots\}$, *then, for any* $u \in H^{n-1}_{p,\theta}$,

$$||S_1^{(m)}u||_{H^n_{q,\sigma}} \leq N||u||_{H^{n-1}_{p,\theta}};$$

(ii) *if* $\tau \leq 0$ *and* $n \in \{2,3,\ldots\}$, *then, for any* $u \in H^{n-2}_{p,\theta}$,

$$||S_1^{(m)}u||_{H^n_{q,\sigma}} \leq N||u||_{H^{n-2}_{p,\theta}},$$

where N *is independent of* u.

Proof. (i) For $k = 0,1,\ldots,n$, we have to estimate the $L_{q,\sigma}$-norm of $M^kD^kS_1^{(m)}u$ through $||u||_{H^{n-1}_{p,\theta}}$. For $k \leq n-1$, we have the desired estimate from Lemma 2.3. Formula (2.11) shows that we need only prove that

$$||M^{n-1}D^nMS_1^{(m)}u||_{L_{q,\sigma}} \leq N||u||_{H^{n-1}_{p,\theta}}. \tag{2.15}$$

We again use the functions ζ and η from the proof of Lemma 2.3 and rewrite (2.13) as follows:

$$\begin{aligned} M^{n-1}D^nMS_1^{(m)}u =& M^{n-1}D^{m+1}T_1D^{n-1}(M^{-1}\bar{\zeta}\bar{u}) \\ &+ M^{n-1}D^{n+m}\tilde{T}_1M^{-1}\eta u. \end{aligned}$$

Then an obvious modification of the argument ending the proof of Lemma 2.3 immediately leads to (2.15).

(ii) As in (i), but now due to (i) instead of Lemma 2.3, we need only prove that

$$||M^{n-1}D^nMS_1^{(m)}u||_{L_{q,\sigma}} \leq N||u||_{H^{n-2}_{p,\theta}}. \tag{2.16}$$

This time we rewrite (2.13) as follows:

$$\begin{aligned} M^{n-1}D^nMS_1^{(m)}u =& M^{n-1}D^{2+m}T_1D^{n-2}(M^{-1}\bar{\zeta}\bar{u}) \\ &+ M^{n-1}D^{n+m}\tilde{T}_1M^{-1}\eta u \end{aligned}$$

and get (2.16) in the same way as above. The lemma is proved. □

The following theorem is the main result of this section. It shows that the solutions of the Cauchy problem for (0.1) are "naturally smoother" than the initial data. This result, interesting in its own right, plays a central role in section 3 in proving the solvability of parabolic equations in weighted spaces.

For $t \geq s$ introduce the operator $\tilde{T}_{s,t}$ so that, for $f \in C_0^\infty(\mathsf{R}^d_+)$, $\tilde{T}_{s,t}f$ is the solution of the equation $u_t(t,x) = a^{ij}(t)u_{x^ix^j}(t,x)$ for $t > s$ and $x \in \mathsf{R}^d_+$ satisfying $u(s,x) = f(x)$ with zero boundary condition at $x^1 = 0$. Due to (2.1), one has a very well-known representation of the kernel of $\tilde{T}_{s,t}$ as the difference of certain Gaussian densities. In the following theorem we use the operator D^m from Lemma 2.2.

THEOREM 2.5. *Assume* $a^{1j}(t) \equiv 0$ *for* $j = 2,\ldots,d$. *Let* $1 < p \leq q < \infty$, $\alpha,\theta,\gamma \in \mathsf{R}$,

$$\alpha \leq 2, \quad d - p/q - p\alpha < \theta < d - 1 + p.$$

Then, for $\alpha_+ = \max(0,\alpha)$, *generalized function* v *given on* R^d_+,

$$\hat{\alpha} = \alpha - (1/p - 1/q)d, \quad (\hat{\theta} - d)/q = (\theta - d)/p,$$

we have

$$||M^{\alpha-1}D^m\tilde{T}_{0,t}v||_{H^{\gamma}_{q,\hat{\theta}}} \le Nt^{(\hat{\alpha}-m)/2-1}||Mv||_{H^{\bar{\gamma}}_{p,\theta}}, \tag{2.17}$$

where $\bar{\gamma} := \gamma + \alpha_+ - 2$ *and* N *depends only on* $d, p, q, \alpha, \theta, \gamma, m,$ *and* δ.

Proof. We give the proof in several steps. In Steps 1 through 3 we assume that $a^{ij} \equiv \delta^{ij}$.

Step 1. Denote $u = Mv$, $\sigma = d + (\theta - d)q/p + q\alpha$. Then it is easy to see (2.17) becomes

$$||S_t^{(m)}u||_{H^{\gamma}_{q,\sigma}} \le Nt^{(\hat{\alpha}-m)/2-1}||u||_{H^{\bar{\gamma}}_{p,\theta}}. \tag{2.18}$$

Now, remember that if $u(t,x)$ is a solution of the heat equation, then for any constant $c > 0$ the function $u(c^2t, cx)$ is also a solution of the same equation. It follows that $\tilde{T}_t u(c\,\cdot)(x) = \tilde{T}_{c^2t}u(cx)$,

$$\begin{aligned} S_t[u(c\,\cdot)](x) &= cM^{-1}\tilde{T}_t[(M^{-1}u)(c\,\cdot)](x) \\ &= cM^{-1}[(\tilde{T}_{c^2t}M^{-1}u)(c\,\cdot)](x) = c^2[M^{-1}\tilde{T}_{c^2t}M^{-1}u](cx), \\ S_t u(x) &= c^2 S_{c^2t}[u(c^{-1}\cdot)](cx). \end{aligned}$$

By adding to this the homogeneity property of $H^{\gamma}_{p,\theta}$-norms, taking $c = t^{-1/2}$, and denoting $v(x) = u(t^{1/2}x)$, we get that, if (2.18) is true for $t = 1$, then

$$\begin{aligned} ||S_t^{(m)}u||_{H^{\gamma}_{q,\sigma}} &= t^{-1-m/2}||[S_1^{(m)}v](t^{-1/2}\cdot)||_{H^{\gamma}_{q,\sigma}} \le Nt^{\sigma/(2q)-1-m/2}||S_1 v||_{H^{\gamma}_{q,\sigma}} \\ &\le Nt^{\sigma/(2q)-1-m/2}||u(t^{1/2}\cdot)||_{H^{\bar{\gamma}}_{p,\theta}} = Nt^{\sigma/(2q)-1-m/2-\theta/(2p)}||u||_{H^{\bar{\gamma}}_{p,\theta}}. \end{aligned}$$

Upon noticing that $\sigma/q - \theta/p = \hat{\alpha}$, we see that we need concentrate only on the case $t = 1$ and prove

$$||S_1^{(m)}u||_{H^{\gamma}_{q,\sigma}} \le N||u||_{H^{\bar{\gamma}}_{p,\theta}}. \tag{2.19}$$

Step 2. By Lemma 2.4, inequality (2.19) holds if $\gamma = 2, 3, \ldots$ and $\alpha \le 0$, so that $\bar{\gamma} = \gamma - 2$. By an interpolation theorem (see Corollary 3.3 of [6]), (2.19) holds if $\gamma \ge 2$ and $\alpha \le 0$. We also observe that, by Lemma 2.2, the kernel of $S_1^{(m)}$ is either symmetric or antisymmetric; hence using duality leads us to

$$||S_1^{(m)}u||_{H^{2-\gamma}_{p',\theta'}} \le N||u||_{H^{-\gamma}_{q',\sigma'}} \tag{2.20}$$

if

$$\gamma \ge 2, \quad p' = p/(p-1), \quad q' = q/(q-1),$$

$$\theta'/p' + \theta/p = d, \quad \sigma'/q' + \sigma/q = d, \tag{2.21}$$

$$\sigma = d + (\theta - d)q/p + q\alpha, \quad \alpha \le 0,$$

$$d - p/q - p\alpha < \theta < d - 1 + p.$$

It turns out that (2.20) implies

$$||S_1^{(m)}u||_{H^{\gamma_1}_{q_1,\sigma_1}} \leq N||u||_{H^{\gamma_1-2}_{p_1,\theta_1}} \tag{2.22}$$

whenever $\gamma_1 \leq 0$,

$$q_1 \geq p_1 > 0, \quad \alpha_1 \leq 0,$$

$$\sigma_1 = d + (\theta_1 - d)q_1/p_1 + q_1\alpha_1, \tag{2.23}$$

$$d - p_1/q_1 - p_1\alpha_1 < \theta_1 < d - 1 + p_1.$$

To derive (2.22) from (2.20), take in (2.20)

$$\gamma = 2 - \gamma_1, \quad p = q_1/(q_1 - 1), \quad q = p_1/(p_1 - 1), \quad \alpha = \alpha_1,$$

$$\theta = pd - \sigma_1(p - 1), \quad \sigma = qd - \theta_1(q - 1).$$

Notice that

$$p' = q_1, \quad q' = p_1, \quad q \geq p,$$

and from the second line in (2.21) we infer that

$$\theta' := p'd - \theta p'/p = \sigma_1(p - 1)p'/p = \sigma_1,$$

$$\sigma' := q'd - \sigma q'/q = \theta_1(q - 1)q'/q = \theta_1.$$

This identifies the subscripts in (2.20) and (2.22). Next, from the second line in (2.23) and from the definition of θ we get

$$\begin{aligned}\theta_1 &= (\sigma_1 - d)p_1/q_1 + d - \alpha_1 p_1 = (\sigma_1 - d)q'/p' + d - \alpha q' \\ &= ((pd - \theta)/(p - 1) - d)q'/p' + d - \alpha q' = (pd - \theta)q'/p - q'd/p' + d - \alpha q'.\end{aligned}$$

We can now check that the third line in (2.21) is consistent with our definitions. We have

$$\begin{aligned}\sigma &:= qd - \theta_1(q - 1) = qd - (pd - \theta)q/p + qd/p' - (q - 1)d + \alpha q \\ &= \theta q/p + qd(p - 1)/p - (q - 1)d + \alpha q = d + (\theta - d)q/p + \alpha q.\end{aligned}$$

Finally, to take care of the inequalities imposed on θ in (2.21), notice that

$$\begin{aligned}\theta &:= q_1 d/(q_1 - 1) - (d + (\theta_1 - d)q_1/p_1 + q_1\alpha_1)/(q_1 - 1) \\ &< q_1 d/(q_1 - 1) - (d + (-p_1/q_1 - p_1\alpha_1)q_1/p_1 + q_1\alpha_1)/(q_1 - 1) = d - 1 + p, \\ \theta &> q_1 d/(q_1 - 1) - (d + (-1 + p_1)q_1/p_1 + q_1\alpha_1)/(q_1 - 1) = d - p/q - p\alpha.\end{aligned}$$

Thus, (2.22) holds, indeed, if conditions (2.23) are satisfied and $\gamma_1 \leq 0$. As mentioned in the beginning of Step 2, this is also true if $\gamma_1 \geq 2$. Interpolation in γ_1 shows that (2.22) holds if conditions (2.23) are satisfied without any restriction on γ_1. This proves our theorem for $\alpha \leq 0$.

Step 3. For $0 \le \alpha \le 2$ we again use interpolation. As in Step 2 one can use Lemma 2.3 and the arguments based on interpolation and duality. Then by combining the result with (2.22), we get that

$$||S_1^{(m)}u||_{H^\gamma_{q,\sigma_0}} \le N||u||_{H^\gamma_{p,\theta_0}}, \quad ||S_1^{(m)}u||_{H^\gamma_{q,\sigma_1}} \le N||u||_{H^{\gamma-2}_{p,\theta_1}}$$

whenever

$$\sigma_0 = d + (\theta_0 - d)q/p + 2q, \quad d - p/q - 2p < \theta_0 < d - 1 + p,$$

$$\sigma_1 = d + (\theta_1 - d)q/p, \quad d - p/q < \theta_1 < d - 1 + p.$$

It follows by interpolation that (2.19) holds if we can find $\kappa \in [0,1]$ and θ_i as above so that

$$\gamma + \alpha - 2 = \kappa(\gamma - 2) + (1-\kappa)\gamma,$$

$$\theta = \kappa\theta_1 + (1-\kappa)\theta_0, \quad \sigma = \kappa\sigma_1 + (1-\kappa)\sigma_0. \tag{2.24}$$

Of course, we take $\kappa = 1 - \alpha/2$. With this choice, the last equation in (2.24) follows from the first one. (Remember that σ is defined in the beginning of Step 1.) It remains only to notice that $\kappa\theta_1 + (1-\kappa)\theta_0$ spans the interval $(d - p/q - p\alpha, d - 1 + p)$ as θ_1 and θ_0 run through $(d - p/q, d - 1 + p)$ and $(d - p/q - 2p, d - 1 + p)$, respectively. This proves the theorem if $a^{ij} \equiv \delta^{ij}$.

Step 4. To consider general a^{ij} with $a^{1j} \equiv 0$ for $j \ge 2$, we apply Theorem 5.1 from the last section. First, by using an obvious time change, one reduces the general situation to the one with $a^{11} \equiv 1$. Bearing in mind an appropriate dilation with respect to $x^2, \dots, x^d$ we may further assume that $(a^{ij}(t)) \ge (\delta^{ij})$.

In Theorem 5.1 take $H = \mathsf{R}$, $\mathcal{U} = \{(u,0)\}$, where $u = u(s,t,x)$ is an arbitrary function defined on $(\mathsf{R}^2 \cap \{s \le t\}) \times \mathsf{R}^d_+$ such that the "norm"

$$||(u,0)||_{\mathcal{U}} := ||M^{\alpha-1}D^m u(0,t_0,\cdot)||_{H^\gamma_{q,\hat\theta}}$$

is finite, where $t_0 > 0$ is a fixed number. Also, let $\mathcal{F} = \{(f,0)\}$, where f is an arbitrary function on $\mathsf{R} \times \mathsf{R}^d_+$ such that $f(0,\cdot) \in M^{-1}H^{\bar\gamma}_{q,\theta}$, with "norm" in $\mathcal{F}$ defined by

$$||(f,0)||_{\mathcal{F}} := ||Mf(0,\cdot)||_{H^{\bar\gamma}_{q,\theta}}.$$

Define $\mathcal{A} = C_0^\infty(\mathsf{R} \times \mathsf{R}^d_+) \times \{0\}$ and $\mathcal{B} = \mathcal{B}_1 \times \{0\}$, where $\mathcal{B}_1$ is the set of all functions u of $(s,t,x) \in (\mathsf{R}^2 \cap \{s \le t\}) \times \mathsf{R}^d_+$, which are bounded and continuous in (s,t,x) along with each their derivative with respect to x and such that $||(u,0)||_{\mathcal{U}} < \infty$.

Then all the assumptions of Theorem 5.1 are satisfied due to obvious properties of the spaces $H^\gamma_{p,\theta}$ and the above treatment of the case $a^{ij} \equiv \delta^{ij}$. By this theorem, (2.17) holds with $t = t_0$ and $v \in C_0^\infty(\mathsf{R}^d_+)$. This yields the result due to denseness of $C_0^\infty(\mathsf{R}^d_+)$ in spaces $H^\nu_{p,\tau}$. The theorem is proved. □

REMARK 2.6. *Obviously, this theorem also holds for complex and Hilbert-space-valued functions* v.

COROLLARY 2.7. *Under the assumptions of Theorem* 2.5*, let* $a^{ij}(t)$ *be infinitely differentiable in* t. *Then, for any* $t > s$, *we have*

$$||M^{\alpha-1}D_s D^m \tilde T_{s,t} v||_{H^\gamma_{q,\hat\theta}} \le N(t-s)^{(\hat\alpha - m)/2 - 2}||Mv||_{H^{\bar\gamma}_{p,\theta}},$$

with N depending only on $d, p, q, \alpha, \theta, \gamma, m, K$, and δ.

To prove this, it suffices to notice that, for $v \in C_0^\infty(\mathsf{R}^d_+)$,

$$\frac{\partial}{\partial s}\tilde{T}_{s,t}v = -a^{11}(s)\frac{\partial^2}{(\partial x^1)^2}\tilde{T}_{s,t}v - \sum_{i,j\geq 2} a^{ij}(s)\frac{\partial^2}{\partial x^i \partial x^j}\tilde{T}_{s,t}v.$$

COROLLARY 2.8. *Under the assumptions of Theorem* 2.5 *let $\hat{\alpha} > 0$ and $M^{-1}u \in H^{\gamma+\alpha}_{p,\theta}$. Then*

$$||M^{\alpha-1}(\tilde{T}_t - 1)u||_{H^\gamma_{q,\hat{\theta}}} \leq Nt^{\hat{\alpha}/2}||M\Delta u||_{H^{\gamma+\alpha-2}_{p,\theta}}.$$

Indeed, if $u \in C_0^\infty(\mathsf{R}^d_+)$,

$$(\tilde{T}_t - 1)u = \int_0^t \tilde{T}_r \Delta u \, dr.$$

Hence,

$$\begin{aligned}||M^{\alpha-1}(\tilde{T}_t - 1)u||_{H^\gamma_{q,\hat{\theta}}} &\leq \int_0^t ||M^{\alpha-1}\tilde{T}_r\Delta u||_{H^\gamma_{q,\hat{\theta}}}\, dr \\ &\leq N\int_0^t r^{\hat{\alpha}/2-1}\, dr||M\Delta u||_{H^{\gamma+\alpha-2}_{p,\theta}} = Nt^{\hat{\alpha}/2}||M\Delta u||_{H^{\gamma+\alpha-2}_{p,\theta}}.\end{aligned}$$

Our assertion follows since $C_0^\infty(\mathsf{R}^d_+)$ is dense in the spaces $H^\gamma_{p,\theta}$ and $||M\Delta u||_{H^\nu_{p,\theta}} \leq N||M^{-1}u||_{H^{\nu+2}_{p,\theta}}$.

3. Equation (0.1) in R^d_+ in spaces with weights. Recall that the spaces $H^\gamma_{p,\theta}$ and the operators $\tilde{T}_{s,t}$ are introduced in the beginning of section 2 and before Theorem 2.5, respectively. Define

$$\tilde{R}f(t) = \int_{-\infty}^t \tilde{T}_{s,t}f(s)\, ds, \quad \tilde{A} = MD^2\tilde{R}M^{-1}f.$$

Existence and uniqueness results for (0.1) in $(0,T)\times\mathsf{R}^d_+$ are based on the following counterpart of Theorem 1.1. In this section we prove Theorem 3.1 under additional assumption (2.1), postponing consideration of the general case until section 4.

THEOREM 3.1. *Let*

$$p, q \in (1,\infty), \quad \gamma \in \mathsf{R}, \quad d - 1 < \theta < d - 1 + p.$$

Then the operator $\tilde{A}$ is uniquely extendable to a bounded operator acting in $L_q(\mathsf{R}, H^\gamma_{p,\theta})$. If we keep the same notation for the extension, then

$$||MD^2\tilde{R}M^{-1}f||_{L_q(\mathsf{R},H^\gamma_{p,\theta})} \leq N(\delta, d, q, p, \gamma)||f||_{L_q(\mathsf{R},H^\gamma_{p,\theta})}. \tag{3.1}$$

Proof. First we prove (3.1) for $q = p$ and $f \in C_0^\infty(\mathsf{R}\times\mathsf{R}^d)$. Without losing generality we assume that $f(s) = 0$ for $s \leq 0$. Then $\tilde{R}M^{-1}f$ is a classical solution of $Lu = -M^{-1}f$ vanishing for $t \leq 0$ and on $x^1 = 0$. Due to our restriction on θ and Theorem 5.6 of [5] we have that (3.1) indeed holds for $q = p$ and $f \in C_0^\infty(\mathsf{R}\times\mathsf{R}^d)$. By

using the fact that $C_0^\infty(\mathsf{R}\times\mathsf{R}^d)$ is dense in $L_q(\mathsf{R},H^\gamma_{p,\theta})$, we conclude that (3.1) holds for $q=p$ and any $f\in L_q(\mathsf{R},H^\gamma_{p,\theta})$.

Below we also use Theorem 4.1 of [5], which says that $||M\Delta\cdot||_{H^\gamma_{p,\theta}}\sim||MD^2\cdot||_{H^\gamma_{p,\theta}}$ in our range of θ. Therefore, instead of considering $\tilde A$ we may and will prove the theorem for

$$\bar A=M\Delta\tilde RM^{-1}f.$$

Notice that, if $f\in C_0^\infty(\mathsf{R}\times\mathsf{R}^d)$ and t is not in the support of $f(s)$ as a function of s, then $\tilde T_{s,t}M^{-1}f(s)$ is infinitely differentiable in x and

$$\bar Af(t)=M\Delta\tilde RM^{-1}f(t)=\int_{-\infty}^t M\Delta\tilde T_{s,t}M^{-1}f(s)\,ds. \tag{3.2}$$

Moreover, estimates (2.2), Theorem 2.5 (with $q=p$ and $\alpha=m=2$), and the boundedness of $\bar A$ as an operator from $L_p(\mathsf{R},H^\gamma_{p,\theta})$ to $L_p(\mathsf{R},H^\gamma_{p,\theta})$ show that (3.2) holds for almost all t outside the support of f if f is a bounded $H^\gamma_{p,\theta}$-valued function with compact support. In other words, for those t,

$$\bar Af(t)=\int_{\mathsf{R}}K(t,s)f(s)\,ds,$$

where the operator $K(t,s)$ is defined by the formula

$$K(t,s)h=I_{t>s}M\Delta\tilde T_{s,t}M^{-1}h.$$

Observe that by Theorem 2.5, $K(t,s)$ is a bounded operator from $H^\gamma_{p,\theta}$ into itself with norm less than $N|t-s|^{-1}$.

Now we claim that $\bar A$ is a bounded operator from $L_q(\mathsf{R},H^\gamma_{p,\theta})$ to $L_q(\mathsf{R},H^\gamma_{p,\theta})$ for $1<q\le p$. Here we prove this under the additional assumption that (2.1) holds. Clearly, we may assume that the coefficients a^{ij} are infinitely differentiable as long as we can prove that the estimates on $\bar A$ are independent of smoothness of a^{ij}. Then, owing to Theorem 1.4, to prove the claim, it suffices to show that the norm of $D_sK(t,s)$ as an operator in $H^\gamma_{p,\theta}$ is less than $N|t-s|^{-2}$ with N depending only on d,p,q,θ,γ,K, and δ. However, this is just the statement of Corollary 2.7 for $\alpha=m=2$.

Thus, $\bar A$ is a bounded operator in $L_q(\mathsf{R},H^\gamma_{p,\theta})$ for $1<q\le p$. The same is true for $1<p\le q$, which is proved by using duality and the fact that the dual to $H^\gamma_{p,\theta}$ is $H^{-\gamma}_{p',\theta'}$, with $1/p+1/p'=1$ and $\theta/p+\theta'/p'=d$, where θ' runs through $(d-1,d-1+p')$ as θ runs through $(d-1,d-1+p)$. The theorem is proved. □

Now we can investigate the solvability of (0.1) in $(0,T)\times\mathsf{R}^d_+$ in weighted spaces. Denote

$$\mathbb{H}^{\gamma,q}_{p,\theta}(T)=L_q((0,T),H^\gamma_{p,\theta}),\quad \mathbb{H}^{\gamma,q}_{p,\theta}=L_q(\mathsf{R},H^\gamma_{p,\theta}),$$

$$\mathbb{L}^q_{p,\theta}(T)=\mathbb{H}^{0,q}_{p,\theta}(T),\quad \mathbb{L}^q_{p,\theta}=\mathbb{H}^{0,q}_{p,\theta}.$$

Remember that the operator L is introduced in (1.1).

THEOREM 3.2. *Let $p,q\in(1,\infty)$, $T\in(0,\infty)$, $\gamma\in\mathsf{R}$,*

$$d-1<\theta<d-1+p,\quad \varepsilon>0.$$

$Mf \in \mathbb{H}^{\gamma,q}_{p,\theta}(T)$, and $M^{2/q-1-\varepsilon}u_0 \in H^{\gamma+2-2/q+\varepsilon}_{p,\theta}$. Then in $M\mathbb{H}^{\gamma+2,q}_{p,\theta}(T)$ there is a unique solution of (0.1) on $(0,T)$ with initial data u_0. For this solution

$$||M^{-1}u||_{\mathbb{H}^{\gamma+2,q}_{p,\theta}(T)} \leq N_1||MLu||_{\mathbb{H}^{\gamma,q}_{p,\theta}(T)} + N_2||M^{2/q-1-\varepsilon}u(0)||_{H^{\gamma+2-2/q+\varepsilon}_{p,\theta}}, \tag{3.3}$$

where $N_1 = N(d,p,q,\delta,K,\theta,\gamma)$ and $N_2 = N(d,p,q,\delta,K,\theta,\gamma,T)$. In addition, if $q = p$, one can allow $\varepsilon = 0$, and then N_2 is independent of T.

Proof. First we reduce the general situation to the one with $u_0 = 0$. From [7] we know that there is a continuation operator P mapping $u_0 \in M^{1+\varepsilon-2/q}H^{\gamma+2-2/q+\varepsilon}_{p,\theta}$ into a function of $(t,x) \in [0,\infty) \times \mathsf{R}^d_+$ which is weakly continuous in t and satisfies $Pu|_{t=0} = u_0$,

$$||M^{-1}Pu_0||_{\mathbb{H}^{\gamma+2,q}_{p,\theta}(T)} + ||M(Pu_0)_t||_{\mathbb{H}^{\gamma,q}_{p,\theta}(T)} \leq N||M^{2/q-1-\varepsilon}u_0||_{H^{\gamma+2-2/q+\varepsilon}_{p,\theta}}.$$

It is also proved that if $q = p$, then the result holds for $\varepsilon = 0$ with N independent of T. By using (2.2) again, we see that indeed everything is reduced to the case $u_0 = 0$.

Now take $f \in C_0^\infty(\mathsf{R}_+ \times \mathsf{R}^d_+)$ and let $u = \tilde{R}f$. Then u is a classical solution of $Lu = -f$ for $t > 0$, $x^1 > 0$ satisfying $u(0) = 0$ and $u = 0$ on $x^1 = 0$. Owing to Theorem 3.1, (2.2), and the fact that $u(t)$ for $t \in [0,T]$ is independent of the values of $f(s)$ for $s \geq T$, we get (3.3) and, from the equation $Lu = -f$, that $||Mu_t||_{\mathbb{H}^{\gamma,q}_{p,\theta}(T)} \leq N||Mf||_{\mathbb{H}^{\gamma,q}_{p,\theta}(T)}$. In particular, for any $T < \infty$,

$$\sup_{t\leq T}||Mu(t,\cdot)||_{H^\gamma_{p,\theta}} \leq N||Mf||_{\mathbb{H}^{\gamma,q}_{p,\theta}(T)}$$

with N independent of f. It follows that if f is in a bounded set in $M^{-1}\mathbb{H}^{\gamma,q}_{p,\theta}(T)$, then u is in a bounded set in

$$M\mathbb{H}^{\gamma+2,q}_{p,\theta}(T) \cap M^{-1}C([0,T],H^\gamma_{p,\theta}).$$

Using obvious approximations proves our assertion about the existence of solutions.

To prove uniqueness, observe that it suffices to prove the a priori estimate (3.3) for any function $M^{-1}u \in \mathbb{H}^{\gamma+2,q}_{p,\theta}(T)$ which is weakly continuous in t and satisfies $u(0) = 0$ and $MLu \in \mathbb{H}^{\gamma,q}_{p,\theta}(T)$. The latter along with (2.2) implies that $Mu_t \in \mathbb{H}^{\gamma,q}_{p,\theta}(T)$. Any such function u belongs to the class $H^{\gamma+2,q}_{p,\theta}(T)$ introduced in [7], where it is also proved that there exists a sequence of functions $u_n(t,x)$ which are infinitely differentiable in x, have support in $[0,T] \times G_n$, where $G_n = \bar{G}_n \subset \mathsf{R}^d_+$, vanish at $t = 0$, have bounded derivative in t, and

$$||M^{-1}(u-u_n)||_{\mathbb{H}^{\gamma+2,q}_{p,\theta}(T)} + ||M(u-u_n)_t||_{\mathbb{H}^{\gamma,q}_{p,\theta}(T)} \to 0.$$

Then, of course, $u_n = -\tilde{R}f_n$ with $f_n(t) = I_{t\in(0,T)}Lu_n(t)$, and

$$||M(Lu-f_n)||_{\mathbb{H}^{\gamma,q}_{p,\theta}(T)} \leq N(||MD^2(u-u_n)||_{\mathbb{H}^{\gamma,q}_{p,\theta}(T)} + ||M(u-u_n)_t||_{\mathbb{H}^{\gamma,q}_{p,\theta}(T)}) \to 0.$$

It remains only to notice that by the first part of the proof

$$||M^{-1}u_n||_{\mathbb{H}^{\gamma+2,q}_{p,\theta}(T)} \leq N_1||Mf_n||_{\mathbb{H}^{\gamma,q}_{p,\theta}(T)}.$$

The theorem is proved. □

The third proof of Theorem 1.1. As in section 1, we may assume that $\gamma = 0$. Also, obviously we need only to prove (1.2) for $f \in C_0^\infty(\mathsf{R} \times \mathsf{R}^d)$. Now comes the crucial observation that, due to self similarity of (1.2) for $\gamma = 0$, we can assume that $f = 0$ for $x^1 \notin (1,2)$ and prove only that

$$||\zeta D^2 R\zeta f||_{\mathbb{L}_p^q} \le N||f||_{\mathbb{L}_p^q} \tag{3.4}$$

for a function $\zeta \in C_0^\infty(\mathsf{R}_+)$ such that $\zeta(r) = 1$ on $[1,2]$.

Next, for $n > 1$ define $f_n(t,x) = f(t,x^1 - n, x')$, $\zeta_n(r) = \zeta(r-n)$. Then it is easily seen that

$$R\zeta f = \lim_{n\to\infty} v_n \quad \text{with} \quad v_n(t,x) = u_n(t,x^1+n,x'), \quad u_n = \tilde{R}(\zeta_n f_n). \tag{3.5}$$

Upon noticing that on the support of ζ_n it holds that $\zeta_n(r) \sim \zeta_n(r)r/n$, by Theorem 3.1 with $\theta = d$ we have

$$\begin{aligned}
||\zeta D^2 v_n||_{\mathbb{L}_p^q}^q = ||\zeta_n D^2 u_n||_{\mathbb{L}_p^q}^q &\le N\int_{\mathsf{R}} \left(\int_{\mathsf{R}^d} |(x^1/n)D^2 u_n(t,x)|^p\,dx\right)^{q/p} dt \\
&\le Nn^{-q}\int_{\mathsf{R}} \left(\int_{\mathsf{R}^d} |x^1\zeta_n f_n(t,x)|^p\,dx\right)^{q/p} dt \\
&\le N\int_{\mathsf{R}} \left(\int_{\mathsf{R}^d} |f_n(t,x)|^p\,dx\right)^{q/p} = N\int_{\mathsf{R}} \left(\int_{\mathsf{R}^d} |f(t,x)|^p\,dx\right)^{q/p}.
\end{aligned}$$

This along with (3.5) yields (3.4) and brings our third proof to an end.

4. Proof of Theorem 3.1 in the general case. We need two lemmas, in the first of which no restriction on θ is imposed. Remember that $Lu = a^{ij}u_{x^ix^j} - u_t$.

LEMMA 4.1. *Let $p \in (1,\infty)$, $n \in \{1,2,\dots\}$, $\gamma \ge \nu$, $\theta \in \mathsf{R}$, $M^{-1}u \in \mathbb{H}_{p,\theta}^{\nu,np}$, and $Mf \in \mathbb{H}_{p,\theta}^{\gamma-2,np}$. Assume u is a solution of $Lu = f$ in $\mathsf{R}\times\mathsf{R}_+^d$. Then $M^{-1}u \in \mathbb{H}_{p,\theta}^{\gamma,np}$ and*

$$||M^{-1}u||_{\mathbb{H}_{p,\theta}^{\gamma,np}} \le N(||MLu||_{\mathbb{H}_{p,\theta}^{\gamma-2,np}} + ||M^{-1}u||_{\mathbb{H}_{p,\theta}^{\nu,np}}), \tag{4.1}$$

where $N = N(d,n,p,\theta,\gamma,\nu,\delta)$.

Proof. Clearly (4.1) becomes stronger if ν decreases. Therefore, we may assume that $\nu = \gamma - k$, where k is an integer, and bearing in mind an obvious induction, we see that, without loss of generality, we may let $\nu = \gamma - 1$.

Now notice that

$$\begin{aligned}
||M^{-1}u||_{\mathbb{H}_{p,\theta}^{\gamma,np}}^{np} &= \int_{\mathsf{R}} ||M^{-1}u(t)||_{H_{p,\theta}^{\gamma}}^{np}\,dt \le N\int_{\mathsf{R}} ||u(t)||_{H_{p,\theta-p}^{\gamma}}^{np}\,dt \\
&= N\sum_{m_1,\dots,m_n=-\infty}^{\infty} e^{(\theta-p)\bar{m}} \int_{\mathsf{R}} \prod_{i=1}^{n} ||u(t,e^{m_i}\cdot)\zeta||_{H_p^\gamma}^p\,dt
\end{aligned} \tag{4.2}$$

with $\bar{m} := m_1 + \cdots + m_n$. Here

$$\begin{aligned}
||u(t,e^m\cdot)\zeta||_{H_p^\gamma}^p &\le N||\Delta[u(t,e^m\cdot)\zeta]||_{H_p^{\gamma-2}}^p \\
&= ||(1-\Delta)^{\gamma/2-1}\Delta[u(t,e^m\cdot)\zeta]||_{L_p}^p \\
&= e^{m(p\gamma-d)}||(\lambda_m - \Delta)^{\gamma/2-1}\Delta[u(t)\zeta_m]||_{L_p}^p,
\end{aligned}$$

where $\lambda_m = e^{-2m}$ and $\zeta_m(x) = \zeta(e^{-m}x)$. Furthermore, $L(u\zeta_m) = \bar{f}_m$, where,

$$\bar{f}_m = f\zeta_m + 2a^{ij}\zeta_{mx^i}u_{x^j} + ua^{11}\zeta_{mx^1x^1},$$

and similarly to the above computation

$$||(\lambda_m - \Delta)^{\gamma/2-1}\bar{f}_m(t)||^p_{L_p} = e^{-m(p\gamma-2p-d)}||\bar{f}_m(t,e^m\cdot)||^p_{H_p^{\gamma-2}}.$$

Therefore, by Lemma 1.6, which is obviously valid for R in place of $(0,T)$, for any $m_1,\dots,m_n$, we have

$$\int_{\mathsf{R}}\prod_{i=1}^n ||u(t,e^{m_i}\cdot)\zeta||^p_{H_p^\gamma}\,dt$$
$$\leq N\int_{\mathsf{R}}\sum_{i=1}^n e^{2m_ip}||\bar{f}_{m_i}(t,e^{m_i}\cdot)||^p_{H_p^{\gamma-2}}\prod_{j\neq i}||u(t,e^{m_j}\cdot)\zeta||^p_{H_p^\gamma}\,dt.$$

Coming back to (4.2), we conclude

$$||M^{-1}u||^{np}_{\mathbb{H}^{\gamma,np}_{p,\theta}} \leq N\int_{\mathsf{R}} F(t)||u(t)||^{(n-1)p}_{H^\gamma_{p,\theta-p}}\,dt,$$

where

$$F(t) := \sum_{m=-\infty}^{\infty} e^{m(\theta+p)}||\bar{f}_m(t,e^m\cdot)||^p_{H_p^{\gamma-2}}.$$

Next we use (see [5]) that the operator M^β is a bounded operator from $H^\gamma_{p,\theta}$ to $H^\gamma_{p,\theta+\beta p}$ and that $M\nabla$ is a bounded operator from $H^\gamma_{p,\theta}$ to $H^{\gamma-1}_{p,\theta}$. Then we find

$$F(t) \leq N\sum_{m=-\infty}^{\infty} e^{m(\theta+p)}||f(t,e^m\cdot)\zeta||^p_{H_p^{\gamma-2}}$$
$$+N\sum_{m=-\infty}^{\infty} e^{m\theta}||u_x(t,e^m\cdot)\zeta'||^p_{H_p^{\gamma-2}} + N\sum_{m=-\infty}^{\infty} e^{m(\theta-p)}||u(t,e^m\cdot)\zeta''||^p_{H_p^{\gamma-2}}$$
$$\leq N\left(||Mf(t)||^p_{H^{\gamma-2}_{p,\theta}} + ||Mu_x(t)||^p_{H^{\gamma-2}_{p,\theta-p}} + ||M^{-1}u(t)||^p_{H^{\gamma-2}_{p,\theta}}\right)$$
$$\leq N\left(||Mf(t)||^p_{H^{\gamma-2}_{p,\theta}} + ||u(t)||^p_{H^{\gamma-1}_{p,\theta-p}}\right) \leq N\left(||Mf(t)||^p_{H^{\gamma-2}_{p,\theta}} + ||M^{-1}u(t)||^p_{H^{\gamma-1}_{p,\theta}}\right).$$

Thus,

$$||M^{-1}u||^{np}_{\mathbb{H}^{\gamma,np}_{p,\theta}}$$
$$\leq NE\int_{\mathsf{R}}\left(||Mf(t)||^p_{H^{\gamma-2}_{p,\theta}} + ||M^{-1}u(t)||^p_{H^{\gamma-1}_{p,\theta}}\right)||M^{-1}u(t)||^{(n-1)p}_{H^\gamma_{p,\theta}}\,dt,$$

and, to get (4.1) for $\nu = \gamma - 1$, it remains only to use Hölder's inequality. The lemma is proved. □

LEMMA 4.2. *Let $p,q \in (1,\infty)$, $d-1 < \theta < d-1+p$, $M^{-1}u \in \mathbb{H}^{2,q}_{p,\theta}$, and $Mf \in \mathbb{L}^q_{p,\theta}$. Assume u is a solution of $Lu = f$ in $\mathsf{R}\times\mathsf{R}^d_+$. Then*

$$||M^{-1}u||_{\mathbb{L}^q_{p,\theta}} \leq N||MLu||_{\mathbb{L}^q_{p,\theta}}, \tag{4.3}$$

where $N = N(d,p,q,\theta)$.

Proof. Our assertion means that the a priori estimate (4.3) holds for any weakly continuous $u(t)$ defined on R and satisfying $M^{-1}u \in \mathbb{H}^{2,q}_{p,\theta}$, which implies $MD^2u \in \mathbb{L}^q_{p,\theta}$ (see (2.2)), and such that the right-hand side of (4.3) is finite, which adds that $Mu_t \in \mathbb{L}^q_{p,\theta}$. This interpretation allows us to use approximations and shows that we may assume $u \in C_0^\infty(\mathsf{R} \times \mathsf{R}^d_+)$.

Denote $Lu = -f$ and let $\tilde{p}(s,t,x,y)$ be the kernel of $\tilde{T}_{s,t}$. It is well known that $\tilde{p}(s,t,x,y)$ can be written as $\tilde{p}(s,t,x^1,y^1,x'-y')$. Then $u(t,x)$ is written as

$$\int_{-\infty}^t \tilde{T}_{t,s} f(s,x)\,ds = \int_{-\infty}^t \int_{\mathsf{R}^d_+} \tilde{p}(s,t,x^1,y^1,y') f(s,y^1,x'-y')\,dyds.$$

Hence by Minkowski's inequality, for any $x^1 > 0$ and t, the norm $||u(t,x^1,\cdot)||_{L_p(\mathsf{R}^{d-1})} =: \bar{u}(t,x^1)$ is less than

$$\int_{-\infty}^t \int_{\mathsf{R}^d_+} \tilde{p}(s,t,x^1,y^1,y') ||f(s,y^1,\cdot - y')||_{L_p(\mathsf{R}^{d-1})}\,dyds$$

$$= \int_{-\infty}^t \int_{\mathsf{R}_+} \tilde{p}(s,t,x^1,y^1) \tilde{f}(s,y^1)\,dy^1 ds =: \tilde{u}(t,x^1),$$

where

$$\tilde{p}(s,t,x^1,y^1) = \int_{\mathsf{R}^{d-1}} \tilde{p}(s,t,x^1,y^1,y')\,dy',$$

$$\tilde{f}(s,y^1) = ||f(s,y^1,\cdot)||_{L_p(\mathsf{R}^{d-1})}.$$

Obviously $\tilde{p}(s,t,x^1,y^1)$ is the fundamental solution of the equation $a^{11}u_{x^1x^1} - u_t = 0$. Hence, by (2.2) and Theorem 3.1 applied for $d = 1$ and $\sigma = \theta - d + 1$, we get

$$||M^{-1}\tilde{u}||_{\mathbb{L}^q_{p,\sigma}} \le ||M^{-1}\tilde{u}||_{\mathbb{H}^{2,q}_{p,\sigma}} \le N||M\tilde{f}||_{\mathbb{L}^q_{p,\sigma}}.$$

It remains only to notice that

$$||M\tilde{f}||^q_{\mathbb{L}^q_{p,\sigma}} = \int_{\mathsf{R}} \left(\int_{\mathsf{R}_+} (x^1)^{\sigma-1+p} |\tilde{f}(t,x^1)|^p\,dx^1 \right)^{q/p} dt = ||Mf||^q_{\mathbb{L}^q_{p,\theta}},$$

$$||M^{-1}u||_{\mathbb{L}^q_{p,\theta}} = ||M^{-1}\bar{u}||_{\mathbb{L}^q_{p,\sigma}} \le ||M^{-1}\tilde{u}||_{\mathbb{L}^q_{p,\sigma}}.$$

The lemma is proved. □

Now we can finish proving Theorem 3.1 in the general case. Remember that the operator $\tilde{A}$ is introduced before Theorem 3.1. As we had mentioned in the proof of Theorem 3.1 in section 3, if $f \in C_0^\infty(\mathsf{R} \times \mathsf{R}^d_+)$, then $\tilde{R}M^{-1}f$ is a classical solution of $Lu = -M^{-1}f$ vanishing for $x^1 = 0$ and for all t sufficiently large negative. It follows quite easily that $M^{-1}\tilde{R}M^{-1}f \in \mathbb{H}^{\gamma,q}_{p,\theta}$ for all γ and q. Therefore, from (2.2) and Lemmas 4.1 and 4.2 we get that $\tilde{A}$ is a bounded operator in $L_{np}(\mathsf{R}, H^\gamma_{p,\theta})$ for $n = 1,2,\ldots$ and $\gamma \ge 0$. By the Marcinkiewicz interpolation theorem we can replace np with any $q \ge p$.

Next we claim that, for $q \geq p$, $\tilde{A}$ is a bounded operator in $L_q(\mathsf{R}, H^\gamma_{p,\theta})$ for any $\gamma \in \mathsf{R}$. As we have seen before, this is equivalent to saying that $M^{-1}\tilde{R}M^{-1}$ is a bounded operator from $\mathbb{H}^{\gamma,q}_{p,\theta}$ to $\mathbb{H}^{\gamma+2,q}_{p,\theta}$ for any γ.

To prove this property of $\tilde{R}M^{-1}$, we introduce the operator

$$\mathcal{L}_b = M^2\Delta + bMD_1, \quad \mathcal{L} = \mathcal{L}_2.$$

We are going to use Theorem 2.16 of [5], which asserts that $\mathcal{L}$ is a bounded one-to-one operator from $H^\gamma_{p,\theta}$ onto $H^{\gamma-2}_{p,\theta}$ and its inverse is also bounded. We are also using the fact (see [5]) that DM and $M^{-\beta}\mathcal{L}_b M^\beta$ are bounded operators from $H^\gamma_{p,\theta}$ into $H^{\gamma-1}_{p,\theta}$ and $H^\gamma_{p,\theta}$ into $H^{\gamma-2}_{p,\theta}$, respectively. Define

$$\bar{R}M^{-1} := \mathcal{L}\tilde{R}M^{-1}\mathcal{L}^{-1} + 2\tilde{R}M^{-1}S = M(M^{-1}\mathcal{L}M)T\mathcal{L}^{-1} + 2MTS,$$

where $T := M^{-1}\tilde{R}M^{-1}$ and S is a "smoothing" operator defined by

$$\begin{aligned} S &= a^{1j}M(D_jM^{-1}\mathcal{L}_1 + M^{-1}\mathcal{L}_1D_j)\tilde{R}M^{-1}\mathcal{L}^{-1} + MD_1\mathcal{L}^{-1} \\ &= a^{1j}MD_j(M^{-1}\mathcal{L}_1M)T\mathcal{L}^{-1} + a^{1j}\mathcal{L}_1(D_jM)T\mathcal{L}^{-1} + MD_1\mathcal{L}^{-1}. \end{aligned}$$

Fix an integer $n \geq 0$ and assume that $M^{-1}\tilde{R}M^{-1}$ is a bounded operator from $\mathbb{H}^{\gamma,q}_{p,\theta}$ to $\mathbb{H}^{\gamma+2,q}_{p,\theta}$ for any $\gamma \geq -n$. By the above result this is true if $n = 0$. Fix a $\gamma \geq -n$. Then T is a bounded operator from $\mathbb{H}^{\gamma,q}_{p,\theta}$ to $\mathbb{H}^{\gamma+2,q}_{p,\theta}$. This and the abovementioned properties of $M^{-\beta}\mathcal{L}_bM^\beta$, D_1M, and MD_1 imply that the operator S is a bounded operator from $\mathbb{H}^{\gamma-1,q}_{p,\theta}$ to $\mathbb{H}^{\gamma,q}_{p,\theta}$. It follows that $M^{-1}\bar{R}M^{-1}$ is a bounded operator from $\mathbb{H}^{\gamma-1,q}_{p,\theta}$ to $\mathbb{H}^{\gamma+1,q}_{p,\theta}$:

$$||M^{-1}\bar{R}M^{-1}f||_{\mathbb{H}^{\gamma+1,q}_{p,\theta}} \leq N||f||_{\mathbb{H}^{\gamma-1,q}_{p,\theta}}. \tag{4.4}$$

Furthermore, it turns out that

$$\bar{R}M^{-1}f = \tilde{R}M^{-1}f \tag{4.5}$$

if $f \in C_0^\infty(\mathsf{R} \times \mathsf{R}^d_+)$. Indeed, the function $v := \tilde{R}M^{-1}\mathcal{L}^{-1}f$ belongs to $M\mathbb{H}^{\gamma+4,q}_{p,\theta}$ (for any γ) and satisfies

$$Lv = -M^{-1}\mathcal{L}^{-1}f.$$

We apply $\mathcal{L}$ to both parts of this equality and, for $u := \mathcal{L}v \in M\mathbb{H}^{\gamma+2,q}_{p,\theta}$, get that

$$Lu = -M^{-1}f + \bar{f},$$

where

$$\bar{f} = 2a^{1j}D_jM^{-1}\mathcal{L}_1v + 2a^{1j}M^{-1}\mathcal{L}_1D_jv + 2D_1\mathcal{L}^{-1}f.$$

Since $u \in M\mathbb{H}^{\gamma+2,q}_{p,\theta}$, it follows that

$$\mathcal{L}\tilde{R}M^{-1}\mathcal{L}^{-1}f =: u = \tilde{R}M^{-1}f - \tilde{R}\bar{f},$$

which is exactly (4.5). Hence, $\bar{R}M^{-1}$ is indeed an extension of $\tilde{R}M^{-1}$.

Thus, (4.4) means that

$$||M^{-1}\tilde{R}M^{-1}f||_{\mathbb{H}^{\gamma+1,q}_{p,\theta}} \le N||f||_{\mathbb{H}^{\gamma-1,q}_{p,\theta}}.$$

We conclude that the assumption that $M^{-1}\tilde{R}M^{-1}$ is a bounded operator from $\mathbb{H}^{\gamma,q}_{p,\theta}$ to $\mathbb{H}^{\gamma+2,q}_{p,\theta}$ for any $\gamma \ge -n$ leads to the conclusion that $M^{-1}\tilde{R}M^{-1}$ is a bounded operator from $\mathbb{H}^{\gamma,q}_{p,\theta}$ to $\mathbb{H}^{\gamma+2,q}_{p,\theta}$ for any $\gamma \ge -(n+1)$.

It follows by induction that the operator A is a bounded operator in $L_q(\mathsf{R}, H^{\gamma}_{p,\theta})$ for any $\gamma \in \mathsf{R}$ if $q \ge p > 1$. After that, repeating literally the corresponding argument from section 3 brings the proof of Theorem 3.1 to an end.

5. A general theorem on equations with coefficients depending only on time. Here we prove a version of Theorem 2.2 of [3]. Let H be a separable Hilbert space. Let $\mathcal{U}$ be a set of couples of H-valued functions $u = (u_1, u_2)$ of $(s,t,x) \in (\mathsf{R}^2 \cap \{s \le t\}) \times \mathsf{R}^d_+$ and $\mathcal{F}$ be a set of couples of H-valued functions $f = (f_1, f_2)$ of (t,x) defined on $\mathsf{R} \times \mathsf{R}^d_+$. We consider couples of functions in order to be able to treat the Cauchy problem and inhomogeneous equations at the same time. Assume that on $\mathcal{U}$ and $\mathcal{F}$ we are given some finite real-valued functions $||\cdot||_{\mathcal{U}}, ||\cdot||_{\mathcal{F}}$ interpreted as "norms."

Assumption 5.1. (i) The "norms" on $\mathcal{U}$ and $\mathcal{F}$ are tangentially translation invariant: for any continuous R^d-valued function $b(t)$ with $b^1(t) \equiv 0$ defined on R and $u \in \mathcal{U}$ and $f \in \mathcal{F}$, we have $u^b \in \mathcal{U}$, $f^b \in \mathcal{F}$, and

$$||u^b||_{\mathcal{U}} = ||u||_{\mathcal{U}}, \quad ||f^b||_{\mathcal{F}} = ||f||_{\mathcal{F}},$$

where $u^b(s,t,x) = u(s,t,x+b(t))$ and $f^b(t,x) = f(t,x+b(t))$.

Assumption 5.2. The set $\mathcal{F}$ contains $\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2$, where $\mathcal{A}_i$ are some tangentially translation invariant sets of functions $g(t,x)$ with compact support $\subset \mathsf{R} \times \mathsf{R}^d_+$ which are continuous in (t,x) together with each their derivative in x.

Assumption 5.3. The set $\mathcal{U}$ contains the set $\mathcal{B} = \mathcal{B}_1 \times \mathcal{B}_2$, where $\mathcal{B}_i$ are some tangentially translation invariant sets of functions $v(s,t,x)$ which are bounded and continuous in (s,t,x) together with each their derivative with respect to x.

Assumption 5.4. If (Ω, Σ, P) is a probability space and (i) $u(\omega,s,t,x)$ is a bounded function such that $u(\omega,\cdot,\cdot,\cdot) \in \mathcal{B}$ for any $\omega \in \Omega$, (ii) $u(\omega,s,t,x)$ is measurable in ω for any (s,t,x), and (iii) $\mathsf{E}\,||u||_{\mathcal{U}} < \infty$, then $\mathsf{E}\,u \in \mathcal{U}$ and $||\mathsf{E}\,u||_{\mathcal{U}} \le \mathsf{E}\,||u||_{\mathcal{U}}$.

Remember that $\tilde{T}_t$ is the semigroup associated with operator Δ with zero boundary condition on $x^1 = 0$ and $\tilde{T}_{s,t}$ is the operator such that, for $f \in C_0^\infty(\mathsf{R}^d_+)$, $\tilde{T}_{s,t}f$ is the solution of the equation $u_t(t,x) = a^{ij}(t)u_{x^ix^j}(t,x)$ for $t > s$ and $x \in \mathsf{R}^d_+$ satisfying $u(s,x) = f(x)$ with zero boundary condition on $x^1 = 0$.

THEOREM 5.1. *Under the above assumptions define the following operators on* $\mathcal{A}$*:*

$$\begin{aligned} R_0 : f = (f_1, f_2) &\to R_0(f_1, f_2)(s,t,x) \\ &= \left(\tilde{T}_{t-s}f_1(s,\cdot)(x), \int_s^t \tilde{T}_{t-r}f_2(r,\cdot)(x)\,dr\right), \\ R : f = (f_1, f_2) &\to R(f_1, f_2)(s,t,x) \\ &= \left(\tilde{T}_{s,t}f_1(s,\cdot)(x), \int_s^t \tilde{T}_{r,t}f_2(r,\cdot)(x)\,dr\right), \end{aligned}$$

and assume that for any $f \in \mathcal{A}$ *we have* $R_0 f \in \mathcal{B}$ *and*

$$||R_0 f||_{\mathcal{U}} \le N_0||f||_{\mathcal{F}},$$

where N_0 is a finite constant. Let $a(t) := (a^{ij}(t)) \geq (\delta^{ij})$ for any t, $a^{1j} \equiv 0$ for $j \geq 2$, and $a^{11} \equiv 1$. Then, for any $f \in \mathcal{A}$, we have $Rf \in \mathcal{U}$ and

$$||Rf||_{\mathcal{U}} \leq N_0 ||f||_{\mathcal{F}}.$$

Proof. Take a probability space (Ω, Σ, P) carrying a d-dimensional Wiener process w_t defined for all $t \in \mathsf{R}$. On R define the following random process:

$$b(t) = \int_0^t \sqrt{a(r) - 1}\, dw_r.$$

Observe that $b^1(t) = 0$ since $a^{11} = 1$ and $a^{1j} = 0$ if $j \geq 2$. Also take a d-dimensional Wiener process B_t independent of w_t. It is well known that for any $s < t$ the random vectors $\eta_{s,t} := b(t) - b(s) + B_t - B_s$ and

$$\zeta_{s,t} := \left(B_t^1 - B_s^1, \int_s^t \left(\sqrt{a(r)}\right)^{2j}\, dw_r^j, \dots, \int_s^t \left(\sqrt{a(r)}\right)^{dj}\, dw_r^j \right)$$

have the same Gaussian distribution and that, for bounded nonrandom functions h and $x^1 > 0$, we have

$$\tilde{T}_{t-s} h(s,\cdot)(x) = \mathsf{E}\, h(s, x + B_t - B_s) I_{\tau(s,x)>t},$$
$$\tilde{T}_{s,t} h(s,\cdot)(x) = \mathsf{E}\, h(s, x + \zeta_{s,t}) I_{\tau(s,x)>t},$$

where

$$\tau(s,x) = \inf\{t : t \geq s, x^1 + B_t^1 - B_s^1 \leq 0\}.$$

Next, for $f \in \mathcal{A}$ we have

$$\begin{aligned} ||\mathsf{E}\left([R_0(f^{-b})]^b\right)||_{\mathcal{U}} &\leq \mathsf{E}||\, [R_0(f^{-b})]^b||_{\mathcal{U}} = \mathsf{E}\, ||R_0(f^{-b})||_{\mathcal{U}} \\ &\leq N_0 \mathsf{E}\, ||f^{-b}||_{\mathcal{F}} = N_0 \mathsf{E}\, ||f||_{\mathcal{F}} = M||f||_{\mathcal{F}}. \end{aligned}$$

It remains only to check that

$$\mathsf{E}\left([R_0(f^{-b})]^b\right) = Rf. \tag{5.1}$$

However, for any bounded Borel h,

$$\begin{aligned} \mathsf{E}\left([\tilde{T}_{t-s}(h^{-b})(s,\cdot)(x)]^b\right) &= \mathsf{E}\, \tilde{T}_{t-s}(h^{-b})(s,\cdot)(x + b(t)) \\ &= \mathsf{E}\, h(s, x + b(t) - b(s) + B_t - B_s) I_{\tau(s,x)>t} \\ &= \mathsf{E}\, h(s, x + \eta_{s,t}) I_{\tau(s,x)>t} = \mathsf{E}\, h(s, x + \zeta_{s,t}) I_{\tau(s,x)>t}, \end{aligned}$$

and the last expression coincides with $\tilde{T}_{s,t} h(s,\cdot)(x)$. This certainly proves (5.1) and with it the theorem. □

Acknowledgments. The author is sincerely grateful to W. Littman for discussing certain issues related to this article and to the referee and I. Gyöngy for several useful comments.

REFERENCES

[1] A. Benedek, A.P. Calderón, and R. Panzone, *Convolution operators on Banach space valued functions*, Proc. Natl. Acad. Sci. USA, 48 (1962), pp. 356–365.

[2] M.A. Krasnoselskii, E.I. Pustylnik, P.E. Sobolevski, and P.P. Zabrejko, *Integral Operators in Spaces of Summable Functions*, Nauka, Moscow, 1966 (in Russian); Noordhoff International Publishing, Leiden, The Netherlands, 1976 (in English).

[3] N.V. Krylov, *A parabolic Littlewood-Paley inequality with applications to parabolic equations*, Topol. Methods Nonlinear Anal., 4 (1994), pp. 355–364.

[4] N.V. Krylov, *An analytic approach to SPDEs*, in Stochastic Partial Differential Equations: Six Perspectives, Math. Surveys Monogr. 64, AMS, Providence, RI, 1999, pp. 185–242.

[5] N.V. Krylov, *Weighted Sobolev spaces and Laplace's equation and the heat equations in a half space*, Comm. Partial Differential Equations, 24 (1999), pp. 1611–1653.

[6] N.V. Krylov, *Some properties of weighted Sobolev spaces in* R^d_+, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 28 (1999), pp. 675–693.

[7] N.V. Krylov, *Some properties of traces for stochastic and deterministic parabolic weighted Sobolev spaces*, J. Funct. Anal., to appear.

[8] O.A. Ladyzhenskaya, V.A. Solonnikov, and N.N. Ural'tceva, *Linear and Quasi-Linear Parabolic Equations*, Nauka, Moscow, 1967 (in Russian); AMS, Providence, RI, 1968 (in English).

[9] W. Littman, C. McCarthy, and N. Riviere, *L^p-multiplier theorems*, Studia Math., 30 (1968), pp. 193–217.

[10] E.M. Stein, *Harmonic Analysis: Real-Variable Methods, Orthogonality and Oscillatory Integrals*, Princeton University Press, Princeton, NJ, 1993.

[11] D.W. Stroock and S.R.S. Varadhan, *Multidimensional Diffusion Processes*, Springer-Verlag, Berlin, New York, 1979.

[12] P. Weidemaier, *On the sharp initial trace of functions with derivatives in* $L_p(0,T;L_p(\Omega))$, Boll. Un. Mat. Ital. B (7), 9 (1995), pp. 321–338.

SIAM J. MATH. ANAL.
Vol. 32, No. 5, pp. 1142–1168

SEQUENTIAL BUCKLING: A VARIATIONAL ANALYSIS*

MARK A. PELETIER†

Abstract. We examine a variational problem from elastic stability theory: a thin elastic strut on an elastic foundation. The strut has infinite length, and its lateral deflection is represented by $u : \mathbb{R} \to \mathbb{R}$. Deformation takes place under conditions of prescribed total shortening, leading to the variational problem

$$\inf \left\{ \frac{1}{2} \int u''^2 + \int F(u) : \frac{1}{2} \int u'^2 = \lambda \right\}. \tag{0.1}$$

Solutions of this minimization problem solve the Euler–Lagrange equation

$$u'''' + pu'' + F'(u) = 0, \qquad -\infty < x < \infty. \tag{0.2}$$

The foundation has a nonlinear stress-strain relationship F', combining a *destiffening* character for small deformation with subsequent *stiffening* for large deformation. We prove that for every value of the shortening $\lambda > 0$ the minimization problem has at least one solution. In the limit $\lambda \to \infty$ these solutions converge on bounded intervals to a periodic profile that is characterized by a related variational problem.

We also examine the relationship with a bifurcation branch of solutions of (0.2), and show numerically that all minimizers of (0.1) lie on this branch This information provides an interesting insight into the structure of the solution set of (0.1).

Key words. fourth-order, Swift–Hohenberg equation, extended Fisher–Kolmogorov equation, localization, localized buckling, concentration-compactness, destiffening, restiffening, destabilization, restabilization

AMS subject classifications. 34C11, 34C25, 34C37, 49N99, 49R99, 73C50, 73H05, 73H10, 73K05, 73K20, 73N20, 73Q05, 73V25, 86A60

PII. S0036141099359925

1. Introduction.

1.1. Localized buckling. Long elastic structures that are loaded in the longitudinal direction can buckle in a localized manner. By this we mean that the lateral deflection is concentrated on a small section of the total length of the structure. A well-known example of this localization phenomenon is the axially loaded cylinder, which buckles in a localized diamond-like pattern [28, 14, 8]. Another example, one which will be the subject of this paper, is the strut on a foundation: a thin elastic layer confined laterally by a different elastic material.

One area of application in which the model of a strut on a foundation has received extensive attention is that of structural geology. In this context the strut represents a thin layer of rock that is embedded in a different type of rock, and the longitudinal compression is the result, directly or indirectly, of tectonic plate movement. In the geological context the most common constitutive assumptions are those of viscous, or visco-elastic, materials; however, there is a case to be made for the importance of elastic effects in the deformation process [21, p. 302], and this is the situation we consider here. An introduction to this field can be found in [21, Chapters 10–15].

*Received by the editors December 2, 1999; accepted for publication (in revised form) September 12, 2000; published electronically February 21, 2001.
http://www.siam.org/journals/sima/32-5/35992.html
†Centrum voor Wiskunde en Informatica, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands (Mark.Peletier@cwi.nl).

Observed geological folds commonly display a certain degree of periodicity. Much of the initial work in this area, initiated by Biot in the late 1950's [1], centered on using the observed period to determine—by doing a parameter fit on the strut model—some of the material properties involved. In the 1970s, with the coming of powerful computational techniques, a consensus arose that folds can be formed in a sequential manner, as depicted by Figure 1.1 [5, 6]. The fold initiates around an imperfection, and as the applied shortening increases, the initial folds lock up and cease to grow, while new folds spawn at neighboring locations. At a given time the resulting profile shows a periodic section flanked by decaying tails; as the shortening increases the periodic section widens. Similar examples of localization followed by spreading are found in axially loaded cylinders [14, 8], in sandwich structures [9], and in kink banding in layered materials [10]. The survey paper [9] discusses these examples from a common perspective.

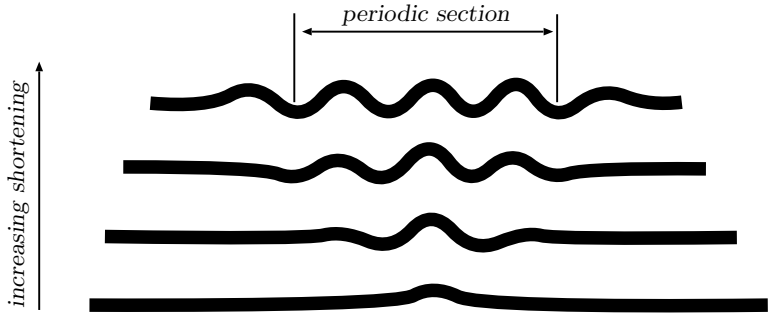


FIG. 1.1. *Folds can form in a sequential manner, driven by increasing shortening (schematic).*

1.2. The modelling. In this paper we investigate the issues of localization and subsequent spreading of deformation for a model of an elastic strut confined by an elastic foundation. We will make a number of important simplifications, and therefore we now discuss the derivation of the equations in some detail.

Our starting point is a thin Euler strut (a strut whose cross-sections remain planar and orthogonal to the center line) of infinite length. Throughout the paper we assume a two-dimensional setting. The independent variable x measures arc length, and we characterize the configuration of the strut by the center-line angle $\theta = \theta(x)$. The strain energy associated with the bending of the strut is equal to $(EI/2)\int \theta'^2(x)dx$. E is Young's modulus and I is the moment of inertia of the cross-section.

The strut is assumed to rest on a foundation of Winkler type, as shown in Figure 1.2. The force response q of this foundation is a function of the local vertical displacement $u(x)$ only, i.e., $q(x) = f(u(x))$. Because of the local character of this response, the strain energy associated with the foundation is equal to $\int F(u(x))dx$, where $F' = f$, $F(0) = 0$. The vertical displacement u and the angle θ are related by $u'(x) = \sin\theta(x)$.

After nondimensionalization the total strain energy for the strut and its foundation is therefore given by

$$\mathcal{W}(\theta) = \frac{1}{2}\int_{-\infty}^{\infty} \theta'^2(x)\,dx + \int_{-\infty}^{\infty} F(u(x))\,dx,$$

where it is understood that $u' = \sin\theta$, $u(-\infty) = 0$. We also define the shortening

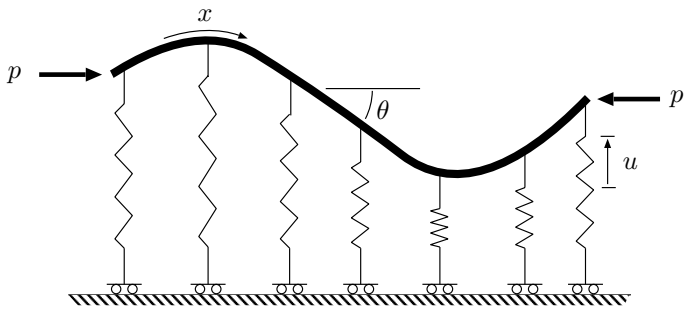


FIG. 1.2. *A strut on an elastic Winkler foundation.*

of the strut, the amount the end-points approach each other because of the deformation $\theta(\cdot)$:

$$\mathcal{J}(\theta) := \int_{-\infty}^{\infty} (1 - \cos\theta(x))\, dx.$$

In engineering it is common to differentiate between dead and rigid loading. In dead loading the external force acting on the structure (in Figure 1.2 the in-plane load p) is prescribed ("controlled" is the usual word, reflecting the possibility of a varying load). In rigid loading a load is applied, but the controlled parameter is the displacement (or some other measure of the deformation). Here the load plays the role of an implied quantity. The two forms of loading share the same equilibria, but the stability properties of these equilibria depend on the form of loading: as a general rule, localized buckles are unstable under dead loading, and stable under rigid loading. (An example of dead loading from daily life is a human being standing on a beer can. As soon as the buckle appears the can collapses completely, showing the instability of the localized buckle under dead loading. However, under rigid loading conditions a variety of localized buckles are witnessed [28]).

With this in mind, we minimize the strain energy $\mathcal{W}$ under a prescribed value λ of the total shortening, i.e., under the condition $\mathcal{J} = \lambda$. While this is a well-posed problem, and one that we intend to return to in subsequent publications, the nonlinearities present render the analysis difficult. We therefore consider a partial linearization of this problem instead, by assuming that u' is small and replacing

$$\theta' = \frac{u''}{\sqrt{1 - u'^2}} \qquad \text{by} \qquad u'' \tag{1.1}$$

and

$$1 - \cos\theta = 1 - \sqrt{1 - u'^2} \qquad \text{by} \qquad \frac{1}{2}u'^2. \tag{1.2}$$

(Note that in doing so we eliminate nonlinearities of a geometrical nature, but retain the nonlinearity in the function F, which is more of a material kind. We discuss this issue further in section 7.) The resulting problem, the central problem in this paper, is

Find a function $u \in H^2(\mathbb{R})$ that solves the minimization problem

$$\inf\{W(u) : J(u) = \lambda\}, \tag{1.3a}$$

where the strain energy W and total shortening J are given by

$$W(u) = \frac{1}{2}\int u'^2 + \int F(u) \qquad \text{and} \qquad J(u) = \frac{1}{2}\int u'^2. \tag{1.3b}$$

A solution u satisfies the Euler–Lagrange equation

$$W'(u) - pJ'(u) = 0, \tag{1.4}$$

for some $p \in \mathbb{R}$, where primes denote Fréchet derivatives, which is equivalent to

$$u'''' + pu'' + f(u) = 0 \qquad \text{on } \mathbb{R}. \tag{1.5}$$

The Lagrange multiplier p is physically interpreted as the in-plane load that is required to enforce the prescribed amount of shortening. Without this load, i.e., when minimizing W without constraint, the sole minimizer would be the trivial state $u \equiv 0$.

Equation (1.5), for various forms of the nonlinearity f, has a history too lengthy to discuss in detail here. Suffice it to mention that it is known, among other names, as the stationary Swift–Hohenberg equation or the stationary extended Fisher–Kolmogorov equation, and that it appears in a host of different applications. We refer the interested reader to the survey articles [2, 3, 18].

1.3. The nonlinearity *F*. The results of this paper depend in a very sensitive manner on the properties of F. In order to describe this we introduce some terminology. Recall that F itself is the potential energy associated with the foundation springs, $F'(u) = f(u)$ is the force associated with a deflection u, and F'' is the *marginal stiffness*.

In the engineering literature *destiffening* refers to a decrease in marginal stiffness, or in everyday language, a weakening of the material. For this model, destiffening refers to a decrease of $F''(u)$ as u moves away from zero (in either positive or negative direction).

The opposite of destiffening is stiffening, which applies to an increase in marginal stiffness as $|u|$ moves away from zero. Although we briefly dwell on such functions in the next section, a more interesting property is what we call *de/restiffening*, or *restiffening* for short: $F''(u)$ decreases for small $|u|$ and becomes increasing for large $|u|$. Throughout this paper we assume a fixed function of restiffening type:

$$F(u) = \frac{1}{2}u^2 - \frac{1}{4}u^4 + \frac{\alpha}{6}u^6, \qquad \alpha \geq \frac{1}{4}. \tag{1.6}$$

Besides the restiffening property this function also has some other desirable qualities, such as

- F is even;
- $F(u) > 0$ if $u \neq 0$;
- $uF'(u) \geq 0$.

We will return to these issues in section 7, where we discuss in some detail the relationship between the results and the function F.

1.4. Results. In this paper we bring together a number of results concerning the minimization problem (1.3), (1.6).

The existence of solutions of the minimization problem (1.3) is not immediate, since the domain is unbounded and therefore minimizing sequences need not be compact. The nonlinearity F is crucial to this issue. To illustrate this, we mention that in the next section we show that a stiffening function F leads to nonexistence:

if $2F(u)/u^2 > F''(0)$ for all $u \neq 0$, then minimizing sequences are never compact, and the infimum is not achieved.

In the parlance of the beginning of this paper, minimizing sequences *delocalize* and spread out. In section 2 we show how the restiffening property of (1.6) guarantees the existence of a minimizer.

The role of λ in problem (1.3) is that of a pure parameter: properties of problem (1.3) for one value of λ are completely decoupled from those for a different value. In addition, minimizers need not be unique. If we choose a minimizer for each value of λ, and denote it by u_λ, then these observations imply that the map $\lambda \mapsto u_\lambda$ may have no continuity properties whatsoever.

In fact, however, the situation is different. The numerical results in section 5 indicate that there is a strong evolutionary aspect, in that the map $\lambda \mapsto u_\lambda$ is "mostly" continuous. In addition, we prove in section 3 that the evolution suggested by Figure 1.1 is essentially correct:

THEOREM. *For any sequence $\lambda_n \to \infty$, a subsequence $u_{\lambda_{n'}}$ converges, after an appropriate translation, to a periodic function $u_\#$. This convergence is uniform on bounded sets.*

The periodic function $u_\#$ solves a related variational problem (see section 3). In section 4 we discuss some symmetry properties of this function.

In section 5 we introduce a numerical method to search for minimizers of (1.3), based on a constrained gradient flow. Figure 1.3 shows some of the results of this calculation. While the form of this curve is unusual at first sight, in section 6 we present an interpretation of this curve in terms of a bifurcation diagram of a related problem ((1.5) for prescribed p). This interpretation, while nonrigorous, gives a satisfactory explanation and raises a few interesting questions as well. We conclude, in section 7, with some comments on the choice of the nonlinearity F.

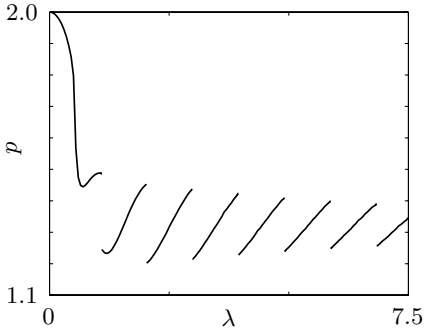


FIG. 1.3. *Plot of the load p_λ associated with a minimizer against λ.*

2. Existence of minimizers. The existence of minimizers of problem (1.3) is a nontrivial problem because of the potential lack of compactness on the unbounded domain $\mathbb{R}$. To illustrate this we consider the case of a completely stiffening function F, one for which $F''(u) > F''(0)$ if $u \neq 0$, or slightly more generally, one for which $2F(u)/u^2 > F''(0)$ if $u \neq 0$. We then have for any $u \in H^2(\mathbb{R})$,

$$W(u) = \frac{1}{2}\int u''^2 + \int F(u) > \frac{1}{2}\int u''^2 + \frac{1}{2}\int u^2 \geq 2J(u), \tag{2.1}$$

where the final inequality follows from the observation that

$$\int u'^2 = -\int u''u \le \frac{1}{2}\int u''^2 + \frac{1}{2}\int u^2.$$

We infer that for any $\lambda > 0$, we have $\inf\{W(u) : J(u) = \lambda\} \ge 2\lambda$, and for any given u this inequality is strict: $W(u) > 2J(u)$.

We now construct an explicit minimizing sequence u_n of problem (1.3) for this potential F. Set

$$u_n(x) = a_n e^{-x^2/n} \sin x, \qquad x \in \mathbb{R}, \tag{2.2}$$

where $a_n \in \mathbb{R}$ is chosen so that $J(u_n) = \lambda$. Note that $a_n \to 0$ as $n \to \infty$. An explicit calculation shows that $W(u_n) \to 2\lambda$; therefore, with the remarks made above in mind, we conclude that $\inf\{W(u) : J(u) = \lambda\} = 2\lambda$ and that the infimum is not attained.

Contained in the argument above is a snippet of information that we will use several times in the proofs that follow. For easy reference we make it a lemma.

LEMMA 2.1. *Let $I \subset \mathbb{R}$ be an interval, bounded or otherwise, and let $u \in H^2(I)$ be such that $uu' = 0$ on ∂I. Then*

$$2\int_I u'^2 \le \int_I u''^2 + \int_I u^2.$$

As above, the proof follows by partial integration.

The theorem below shows that, in contrast to the example above, the infimum is attained if F is not of completely stiffening type, but has a destiffening character for small u (i.e., $F''(u) < F''(0)$ for small $u \ne 0$). We can interpret the situation in the following way. A destiffening quality ($F'' < F''(0)$) favors localized deformation, therefore causing minimizing sequences to be compact, resulting in the existence of minimizers on unbounded domains. A stiffening potential favors delocalization, spreading, of the deformation, as illustrated by the minimizing sequence (2.2). If the two characters are combined, as in the potential (1.6), then the destiffening character for small u is sufficient to guarantee the existence of minimizers, regardless of the behavior for large u. On the other hand, the restiffening character in F becomes noticeable for larger values of λ, in which an equilibrium between localizing and spreading effects creates a periodic structure. We will return to this issue in the next section.

Note that on a bounded interval, given appropriate boundary conditions, a minimizer always exists. One might wonder whether the problem would not be simplified by working on a bounded interval instead of on $\mathbb{R}$. In fact, we expect a strong correspondence between the (non-)existence of minimizers on $\mathbb{R}$ and the form of the minimizers on large but bounded intervals: if existence holds on $\mathbb{R}$, then minimizers on intervals will be localized and largely independent of the size of the interval; but if there is nonexistence on $\mathbb{R}$, then minimizers on the interval will be spread out, with a small amplitude, similar to the sequence u_n above. (See [7] for a discussion of the purely stiffening nonlinearity on a bounded interval). From the point of view of the developments later in this paper, the current problem, with a restiffening foundation, is fundamentally different from the purely stiffening case. In addition, we will use the unbounded domain in the convergence result of Theorem 3.1 and in the comparison with a bifurcation diagram on $\mathbb{R}$ in section 6. With this in mind we choose to consider the problem on the unbounded domain $\mathbb{R}$.

Throughout this paper we define

$$W_\lambda = \inf_{u\in C_\lambda} W(u), \qquad C_\lambda = \{u \in H^2(\mathbb{R}) : J(u) = \lambda\}.$$

THEOREM 2.2. *Let F be as given in* (1.6). *Then for each $\lambda > 0$ there exists $u \in C_\lambda$ that minimizes* (1.3).

Before we prove this theorem we derive some auxiliary properties.

LEMMA 2.3. *For all $\lambda > 0$,*

1. *$W_\lambda < 2\lambda$.*
2. *If $u_n \in \mathcal{C}_\lambda$ is a minimizing sequence of W, then*

$$\limsup_{n\to\infty} \|u_n\|_{L^\infty(\mathbb{R})} \le M\lambda$$

for some constant M.

Proof. Define the explicit sequence

$$u_\varepsilon(x) = a_\varepsilon \varepsilon^{1/2} \operatorname{sech}(\varepsilon x)\cos x,$$

where a_ε is chosen such that $J(u_\varepsilon) = \lambda$ (note that $a_\varepsilon = O(1)$ as $\varepsilon \to 0$). This sequence satisfies $W(u_\varepsilon) \to 2\lambda$, implying $W_\lambda \le 2\lambda$. For the strict inequality we compute

$$\begin{aligned}\lambda &= \frac{1}{2}\int u_\varepsilon'(x)^2\,dx \\ &= \frac{1}{2}a_\varepsilon^2\left\{\varepsilon\int \operatorname{sech}^2(\varepsilon x)\sin^2 x\,dx - 2\varepsilon^2\int \operatorname{sech}(\varepsilon x)\operatorname{sech}'(\varepsilon x)\cos x\sin x\,dx\right. \\ &\qquad \left. + \varepsilon^3\int (\operatorname{sech}'(\varepsilon x))^2\cos^2 x\,dx\right\}. \end{aligned} \tag{2.3}$$

Note that

$$\begin{aligned}\int \operatorname{sech}(\varepsilon x)\operatorname{sech}'(\varepsilon x)\cos x\sin x\,dx& \\ &= \frac{1}{4\varepsilon}\int (\operatorname{sech}^2(\varepsilon x))'\sin 2x\,dx = \frac{1}{2\varepsilon}\int \operatorname{sech}^2(\varepsilon x)\cos 2x\,dx \\ &= \frac{1}{\varepsilon^2}\sqrt{\frac{\pi}{2}}\,\widehat{(\operatorname{sech}^2)}\left(\frac{2}{\varepsilon}\right),\end{aligned}$$

where $\hat{}$ denotes the Fourier transform

$$\hat v(\omega) = \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} v(x)e^{-i\omega x}dx.$$

Since $\operatorname{sech}^2 \in \mathcal{S}$, the set of smooth rapidly decreasing functions, we have $\widehat{(\operatorname{sech}^2)} \in \mathcal{S}$, and therefore

$$\int \operatorname{sech}(\varepsilon x)\operatorname{sech}'(\varepsilon x)\cos x\sin x\,dx = o(\varepsilon^k) \qquad \text{for all } k \in \mathbb{N}.$$

Using the same ideas to estimate the first and third terms in (2.3) we find

$$\begin{aligned}\int \operatorname{sech}^2(\varepsilon x)\sin^2 x\,dx &= \frac{1}{2}\int \operatorname{sech}^2(\varepsilon x)\,dx + o(\varepsilon^k), \\ \int (\operatorname{sech}'(\varepsilon x))^2\cos^2 x\,dx &= \frac{1}{2}\int (\operatorname{sech}'(\varepsilon x))^2\,dx + o(\varepsilon^k)\end{aligned}$$

for all $k \in \mathbb{N}$. Consequently (2.3) implies

$$a_\varepsilon^2 = 4\lambda(1 + O(\varepsilon^2)) \left(\int_{\mathbb{R}} \operatorname{sech}^2(y)\, dy \right)^{-1}.$$

Using this we compute

$$\frac{1}{2} \int u_\varepsilon^2 = \lambda(1 + O(\varepsilon^2)), \tag{2.4}$$

$$\int u_\varepsilon^4 = c_1 \varepsilon (1 + O(\varepsilon^2)), \tag{2.5}$$

$$\int u_\varepsilon^6 = c_2 \varepsilon^2 (1 + O(\varepsilon^2)), \tag{2.6}$$

$$\frac{1}{2} \int u_\varepsilon''^2 = \lambda(1 + O(\varepsilon^2)) \tag{2.7}$$

for some constants $c_1, c_2 > 0$. For the last equality above we apply the same argument as for $\int u_\varepsilon'^2$ to eliminate the cross-product terms. Uniting these estimates we conclude that

$$W(u_\varepsilon) = \lambda(2 + O(\varepsilon^2)) - c_1 \varepsilon,$$

and hence

$$\inf_{\mathcal{C}_\lambda} W(u) < 2\lambda.$$

For part 2 we first note that since $\alpha > 3/16$, there exists $\beta > 0$ such that

$$F(u) \geq \frac{\beta}{2} u^2 \qquad \text{for} \qquad u \in \mathbb{R}. \tag{2.8}$$

By part 1 we can restrict our attention to minimizing sequences that satisfy $W(u_n) \leq 2\lambda$; we have

$$\|u_n\|_{L^\infty(\mathbb{R})}^2 \leq C \|u_n\|_{H^1(\mathbb{R})}^2 \leq 2C \max\{1, 1/\beta\}\, W(u_n) \leq 4C\lambda \max\{1, 1/\beta\}. \qquad \square$$

Remark 2.1. The proof of part 1 of the lemma above uses the relative importance of the destiffening quartic term: the destiffening is of order ϵ, while the "noise" associated with the nonconstant amplitude in u_n is of order ϵ^2 (as shown by the estimates (2.4) and (2.7)). It follows that for a destiffening character of higher order, e.g., a function F of the type $u^2/2 - u^8/8 + \alpha u^{10}/10$, this method of proof does not apply, since the destiffening will be dwarfed by the noise. However, numerical tests have shown that for such functions F the minimization problem still admits solutions, and that the assertion of the lemma still holds.

COROLLARY 2.4. *Let u_n be a minimizing sequence for problem* (1.3). *Then*

$$\liminf_{n\to\infty} \|u_n\|_{L^\infty(\mathbb{R})} = m(\lambda) > 0.$$

Proof. If $\|u_n\|_{L^\infty(\mathbb{R})} \to 0$, then

$$\frac{\frac{1}{2}\int u_n''^2 + \frac{1}{2}\int u_n^2}{W(u_n)} \to 1 \qquad \text{as} \quad n \to \infty. \tag{2.9}$$

Since

$$2\lambda = \int {u_n'}^2 = -\int u_n'' u_n \leq \frac{1}{2}\int u_n''^2 + \frac{1}{2}\int u_n^2,$$

we infer from (2.9) that $\liminf W(u_n) \geq 2\lambda$, which contradicts part 1 of Lemma 2.3. □

In addition, we need an a priori result on minimizers, which is proved in the appendix, as shown below.

LEMMA 2.5. *Let $u \in H^2(\mathbb{R})$ be a solution of (1.3). Then $p < 2$.*

We now continue with the proof of the main theorem of this section.

Proof of Theorem 2.2. The proof follows quite closely the outline of the examples given in [12, 13]. Let u_n be a minimizing sequence, and consider $\rho_n = {u_n'}^2/2$, so that $\rho_n \geq 0$ and $\int \rho_n = 1$. Of the three possibilities for this sequence, vanishing, dichotomy, and compactness, we show that neither vanishing nor dichotomy can occur, leaving compactness as the only possibility.

Vanishing cannot occur. Suppose that

$$\sup_x \int_{x-R}^{x+R} {u_n'}^2 \to 0 \qquad \text{for all} \qquad R > 0.$$

We can choose $x = 0$ as the location of a maximum of each $|u_n|$, and by Corollary 2.4 we then have $u_n(0) \geq m(\lambda) > 0$ (changing u_n into $-u_n$ if necessary). Consequently

$$\liminf_{n\to\infty} \int_{\mathbb{R}} F(u_n) \geq \liminf_{n\to\infty} \frac{\beta}{2}\int_{-R}^{R} u_n^2 \geq \beta m(\lambda)^2 R,$$

which is unbounded as $R \to \infty$. This contradicts $\limsup W(u_n) < 2\lambda$.

Dichotomy cannot occur. For any given $\lambda > 0$, dichotomy is contradicted, proving compactness of the minimizing sequence and therefore existence of a minimizer, if

$$W_\lambda < W_{\theta\lambda} + W_{(1-\theta)\lambda} \tag{2.10}$$

for all $\theta \in (0,1)$ (see [12, 13]). We shall show that (2.10) holds for all $\lambda > 0$ and $\theta \in (0,1)$.

Define

$$A = \{\mu > 0 : (1.3) \text{ has a solution for all } 0 < \lambda \leq \mu\}.$$

First we show that A is nonempty. There exist $\bar{u}, \delta > 0$ such that $2F(u) - uf(u) \geq \delta u^4$ for all $|u| \leq \bar{u}$. Choose λ_0 small enough to ensure that $2M\lambda \leq \bar{u}$ for all $0 < \lambda < \lambda_0$, and pick $0 < \lambda < \lambda_0$. Let v_n be a minimizing sequence such that $J(v_n) = \lambda$; without loss of generality we suppose that $\|v_n\|_{L^\infty(\mathbb{R})} \leq \bar{u}$. Then

$$\begin{aligned}
\frac{d}{d\mu}\frac{W(\mu v_n)}{J(\mu v_n)}\bigg|_{\mu=1} &= \frac{1}{J(v_n)^2}\left(J(v_n)W'(v_n)v_n - W(v_n)J'(v_n)v_n\right) \\
&= \frac{1}{J(v_n)}\left(\int v_n f(v_n) - 2\int F(v_n)\right) \\
&\leq -\frac{\delta}{\lambda}\int v_n^4. \qquad (2.11)
\end{aligned}$$

Since $\|v_n'\|^2_{L^2(\mathbb{R})} = 2\lambda$ is bounded and $\|v_n\|_{L^\infty(\mathbb{R})} > m(\lambda)$, the last term above is bounded away from zero as $n \to \infty$. Therefore W_λ/λ is a strictly decreasing function of λ for $0 < \lambda < \lambda_0$; this shows that $(0, \lambda_0) \subset A$, since we have for any $\theta \in (0,1)$

$$\begin{aligned} W_\lambda = \lambda \frac{W_\lambda}{\lambda} &< \lambda \left(\theta \frac{W_{\theta\lambda}}{\theta\lambda} + (1-\theta) \frac{W_{(1-\theta)\lambda}}{(1-\theta)\lambda} \right) \\ &= W_{\theta\lambda} + W_{(1-\theta)\lambda}. \end{aligned}$$

To show that A is open, suppose that there exists a sequence $\lambda_n + \varepsilon_n \notin A$, $\lambda_n + \varepsilon_n \downarrow \lambda$, $\lambda \in A$, and $\varepsilon_n \to 0$, such that

$$W_{\lambda_n + \varepsilon_n} = W_{\lambda_n} + W_{\varepsilon_n}.$$

Since $\liminf W_{\varepsilon_n}/\varepsilon_n = 2$—by an argument similar to that of Corollary 2.4—this implies

$$\limsup_{\varepsilon \to 0} \frac{1}{\varepsilon} (W_{\lambda+\varepsilon} - W_\lambda) \geq 2. \tag{2.12}$$

However, since $\lambda \in A$, there exists $u \in \mathcal{C}_\lambda$ with $W(u) = W_\lambda$, and by Lemma 2.5 the associated load satisfies $p < 2$. Then

$$\frac{d}{d_\mu} W(\mu u)\bigg|_{\mu=1} = W'(u)u = pJ'(u)u = p \, \frac{d}{d_\mu} J(\mu u)\bigg|_{\mu=1}.$$

Consequently

$$\limsup_{\varepsilon \to 0} \frac{1}{\varepsilon} (W_{\lambda+\varepsilon} - W_\lambda) \leq p < 2,$$

contradicting (2.12).

Finally, we show that A is closed by the following claim: If u and v are minimizers of W at the respective values of λ, then

$$\inf\{W(z) : J(z) = J(u) + J(v)\} < W(u) + W(v). \tag{2.13}$$

This proves that A is closed by the following argument. Suppose $A \supset (0, \lambda_0)$; for all $\theta \in (0,1)$ we have functions u and v that minimize W under the constraints $J(u) = \theta\lambda_0$ and $J(v) = (1-\theta)\lambda_0$. Inequality (2.13) then reduces to (2.10), implying that problem (1.3) also has a solution for $\lambda = \lambda_0$.

To prove (2.13) we choose two sequences $x_n \geq 0$, $y_n \leq 0$, with $x_n \to \infty$ and $y_n \to -\infty$ with certain properties detailed below. We introduce a notation for integrals over a part of $\mathbb{R}$:

$$W_{[a,b]}(u) = \frac{1}{2} \int_a^b u''^2 + \int_a^b F(u) \qquad \text{and} \qquad J_{[a,b]}(u) = \frac{1}{2} \int_a^b u'^2.$$

Setting $p = \max\{p_u, p_v\}$ (the maximum of the two values of the load associated with u and v) we require of the sequences x_n, y_n that there exists an $\varepsilon > 0$ such that

$$p \leq \min \left\{ \frac{W_{[x_n,\infty)}(u)}{J_{[x_n,\infty)}(u)}, \frac{W_{(-\infty,y_n]}(v)}{J_{(-\infty,y_n]}(v)} \right\} - \varepsilon \qquad \text{for all } n.$$

This is possible since u and v are small at infinity, and therefore

$$\limsup W_{[x_n,\infty)}(u)/J_{[x_n,\infty)}(u) \geq 2.$$

In addition we assume that $(u,u')(x_n) = (v,v')(y_n)$ for all n. This is also possible since for large K the set $\{(u,u')(x) : x > K\} \subset \mathbb{R}^2$ is a spiral around the origin. The same is true for $\{(v,v')(x) : x < -K\}$ but for v the spiral rotates in the opposite direction. It follows that there is a countably infinite set of intersections of the two spirals, corresponding to pairs (x_n, y_n) with $x_n \to \infty$, $y_n \to -\infty$.

Now pick $\hat{u}, \hat{v} \in H^2(\mathbb{R})$ such that $\operatorname{supp}\hat{u} \subset (-\infty, 0)$ and $\operatorname{supp}\hat{v} \subset (0,\infty)$ and that in addition $J'(u)\hat{u} = J'(v)\hat{v} = 1$. Define

$$z_n(x) = \begin{cases} u(x_n + x) + \gamma_n \hat{u}(x_n + x) & \text{for } x < 0, \\ v(y_n + x) + \gamma_n \hat{v}(y_n + x) & \text{for } x > 0. \end{cases}$$

Here γ_n is fixed by the requirement $J(z_n) = J(u) + J(v)$:

$$J(z_n) = \frac{1}{2}\int_{-\infty}^{x_n} u'^2 + \gamma_n \int_{-\infty}^{x_n} u'\hat{u}' + \frac{\gamma_n^2}{2}\int_{-\infty}^{x_n} \hat{u}'^2 + \frac{1}{2}\int_{y_n}^{\infty} v'^2 + \gamma_n \int_{y_n}^{\infty} v'\hat{v}' + \frac{\gamma_n^2}{2}\int_{y_n}^{\infty} \hat{v}'^2.$$

Since $\operatorname{supp}\hat{u} \cap [x_n, \infty) = \emptyset$,

$$\gamma_n \int_{-\infty}^{x_n} u'\hat{u}' = \gamma_n \int_{-\infty}^{\infty} u'\hat{u}' = \gamma_n J'(u)\hat{u} = \gamma_n,$$

so that

$$\begin{aligned} J(z_n) &= \frac{1}{2}\int_{-\infty}^{x_n} u'^2 + \frac{1}{2}\int_{y_n}^{\infty} v'^2 + 2\gamma_n + C\gamma_n^2 \\ &= J(u) + J(v) - J_{[x_n,\infty)}(u) - J_{(-\infty,y_n]}(v) + 2\gamma_n + C\gamma_n^2, \end{aligned}$$

where $C = (1/2)\int(\hat{u}'^2 + \hat{v}'^2)$. It follows that γ_n satisfies

$$\gamma_n = \frac{1}{2}J_{[x_n,\infty)}(u) + \frac{1}{2}J_{(-\infty,y_n]}(v) - \frac{C}{2}\gamma_n^2$$

as $n \to \infty$. Note that $\gamma_n \to 0$.

Putting it all together,

$$\begin{aligned} W(z_n) &= W_{(-\infty,x_n]}(u + \gamma_n\hat{u}) + W_{[y_n,\infty)}(v + \gamma_n\hat{v}) \\ &= W(u + \gamma_n\hat{u}) + W(v + \gamma_n\hat{v}) - W_{[x_n,\infty)}(u + \gamma_n\hat{u}) - W_{(-\infty,y_n]}(v + \gamma_n\hat{v}) \\ &= W(u) + W(v) + \gamma_n(W'(u)\hat{u} + W'(v)\hat{v}) + O(\gamma_n^2) - W_{[x_n,\infty)}(u) \\ &\qquad - W_{(-\infty,y_n]}(v) \\ &= W(u) + W(v) + \gamma_n(p_u + p_v) + O(\gamma_n^2) - W_{[x_n,\infty)}(u) - W_{(-\infty,y_n]}(v) \\ &\leq W(u) + W(v) + 2\gamma_n p - (p+\varepsilon)(J_{[x_n,\infty)}(u) + J_{(-\infty,y_n]}(v)) + O(\gamma_n^2) \\ &\leq W(u) + W(v) - 2\varepsilon\gamma_n + O(\gamma_n^2). \end{aligned}$$

This last inequality proves the claim (2.13) and therefore Theorem 2.2. ∎

The definition of W_λ provides no explicit continuity properties with respect to variation of λ. However, the variational character can be exploited to derive an interesting semiconvexity property.

LEMMA 2.6. *There exists $C > 0$ such that*

$$\frac{d^2}{d\lambda^2} W_\lambda \le \frac{C}{\lambda} \qquad \text{for all} \qquad \lambda > 0,$$

in the sense of distributions.

Proof. Note that $u^2 f'(u) - uf(u) \le 24F(u)$ for all $u \in \mathbb{R}$. Choose $\lambda > 0$, and let u achieve W_λ. Setting $v_h = u\sqrt{1 + h/\lambda}$, we have $J(v_h) = \lambda + h$ and

$$\begin{aligned} \frac{d^2}{dh^2} W(v_h)\Big|_{h=0} &= \frac{1}{4\lambda^2} \left\{ W''(u) \cdot u \cdot u - W'(u) \cdot u \right\} \\ &= \frac{1}{4\lambda^2} \int (u^2 f'(u) - uf(u)) \\ &\le \frac{6W(u)}{\lambda^2} \le \frac{12}{\lambda}. \end{aligned}$$

This implies the result. ∎

Lemma 2.6 implies that the left and right derivatives of W_λ with respect to λ are well defined. Note that the Euler–Lagrange equation (1.4) implies that if W_λ is achieved at $\lambda = \lambda_0$ by u_{λ_0}, with load p_{λ_0}, then $\partial W_\lambda / \partial \lambda(\lambda_0 -) \ge p_{\lambda_0} \ge \partial W_\lambda / \partial \lambda(\lambda_0 +)$. It follows that any jumps in p_λ must be downward (for increasing λ).

3. Appearance of a periodic section. In the introduction we mentioned the locking-up and spreading of the deformation as the shortening increases. If this process is continued, we expect a periodic section to build up, flanked by spreading tails. The following theorem makes this precise for the model considered in this paper.

THEOREM 3.1. *For any sequence $\lambda_n \to \infty$, a subsequence $u_{\lambda_{n'}}$ converges, after an appropriate translation, to a periodic function $u_\#$. This convergence is in $C^k(K)$ for all $k \ge 0$ and for all compact sets $K \subset \mathbb{R}$. The periodic function $u_\#$ solves the minimization problem*

$$M_\# = \inf \left\{ \frac{W(u)}{J(u)} : u \in H^2_{\mathrm{loc}}(\mathbb{R}) \text{ periodic} \right\}. \tag{3.1}$$

In addition, as $\lambda_{n'} \to \infty$, $p_{\lambda_{n'}} \to M_\#$.

In the formulation of this theorem, as in the rest of this paper, the functionals W and J will be defined on periodic functions $u \in H^2_{\mathrm{loc}}(\mathbb{R})$ by restricting the integrals to a period and normalizing, i.e., if u has period T, then

$$W(u) = \frac{1}{2T} \int_0^T u''^2 + \frac{1}{T} \int_0^T F(u).$$

Before entering the details of the proof, we should briefly comment on the appearance of the new minimization problem (3.1). If u minimizes W/J among all periodic functions, then by choosing periodic test functions $\phi \in H^2_{\mathrm{loc}}(\mathbb{R})$ with the same period and considering the perturbations $u + \varepsilon\phi$ we derive the Euler–Lagrange equation

$$0 = \frac{1}{J(u)} \left\{ W'(u) \cdot \phi - \frac{W(u)}{J(u)} J'(u) \cdot \phi \right\}.$$

Comparison with (1.4) shows that u solves the same ODE as u_λ, and the load is numerically equal to the optimal quotient $W(u)/J(u) = M_\#$.

We conjecture that the solution of (3.1) is unique for the function F that we consider in this paper. However, it is not difficult to construct a different function F for which uniqueness does not hold. (One could construct a function F which is identical to (1.6) over the range of $u_\#$, but is different for (much) larger values of $|u|$. Then $u_\#$ remains a local minimum for the minimization problem (3.1), but an additional minimum may exist with a much larger amplitude. By adjusting F this function can be given the same value of the ratio W/J as $u_\#$).

Proof. The proof falls apart in five steps.

Step 1. $\limsup_{\lambda\to\infty} W_\lambda/\lambda \le M_\#$. Indeed, if v is a periodic function, then $v_\lambda(x) = \eta(|x| - \mu)v(x)$ belongs to C_λ for some $\mu = \mu(\lambda)$. Here η is a smooth cut-off function satisfying

$$\eta(x) = \begin{cases} 1 & x \le 0, \\ 0 & x \ge 1. \end{cases}$$

Then $W(v_\lambda)/\lambda = W(v_\lambda)/J(v_\lambda) \to W(v)/J(v)$ as $\lambda \to \infty$; therefore

$$\limsup_{\lambda\to\infty} \frac{W_\lambda}{\lambda} \le \limsup_{\lambda\to\infty} \frac{W(v_\lambda)}{\lambda} = \frac{W(v)}{J(v)},$$

from which it follows that $\limsup_{\lambda\to\infty} W_\lambda/\lambda \le M_\#$.

Step 2. Translation of u_λ and construction of a periodic function w_λ.

We first note that by the assumption $\alpha \ge 1/4$ the nonlinearity F is increasing in $|u|$. This implies that $p \ge 0$ by the equation (obtained by multiplying (1.5) by u and integrating)

$$p\int u'^2 = \int u''^2 + \int uf(u). \tag{3.2}$$

As a result the origin is a saddle-focus for (1.5) (when viewed as a dynamical system in x), and orbits in the stable and unstable manifold oscillate around zero.

For a given λ we divide $\mathbb{R}$ into intervals $[x_i, x_{i+1})$ delimited by the stationary points x_i of u_λ. Note that the oscillation mentioned above implies that none of the intervals $[x_i, x_{i+1})$ is unbounded. We calculate the ratio r_i of the local values of W and J for each of these intervals,

$$r_i = \frac{\frac12 \int_{x_i}^{x_{i+1}} u_\lambda''^2 + \int_{x_i}^{x_{i+1}} F(u_\lambda)}{\frac12 \int_{x_i}^{x_{i+1}} u_\lambda'^2}.$$

For large $|x|$, $F(u_\lambda) \sim u_\lambda^2/2$, and therefore $\liminf_{i\to\pm\infty} r_i \ge 2$. Since $W(u_\lambda)/\lambda$ is a convex combination of $\{r_i\}$,

$$\frac{W(u_\lambda)}{\lambda} = \sum_{i\in\mathbb{Z}} \frac{r_i}{2\lambda} \int_{x_i}^{x_{i+1}} u_\lambda'^2, \qquad \sum_{i\in\mathbb{Z}} \frac{1}{2\lambda} \int_{x_i}^{x_{i+1}} u_\lambda'^2 = \frac{J(u_\lambda)}{\lambda} = 1,$$

and since $W(u_\lambda)/\lambda < 2$, there exists $i \in \mathbb{Z}$ such that r_i is minimal among all r_i, and for this i we have $r_i < 2$. Fixing i we translate u_λ such that the interval $[x_i, x_{i+1})$ becomes $[0, T/2)$. The periodic function w_λ, with period T, is now defined to be equal

to u_λ on $[0, T/2)$, and to be even around 0 and around $T/2$, as shown in Figure 3.1. Note that by the choice of i we have

$$\frac{W(w_\lambda)}{J(w_\lambda)} = r_i < \frac{W(u_\lambda)}{\lambda}. \tag{3.3}$$

Remark also that this inequality implies that any localized function has a ratio W/J that is strictly larger than $M_\#$. To indicate the dependence of T on λ we write T_λ.

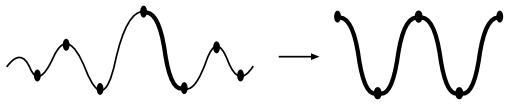


FIG. 3.1. *A section between two stationary points is replicated.*

Step 3. $\lim_{\lambda\to\infty} W(w_\lambda)/J(w_\lambda) = M_\#$. This follows from the sequence of inequalities

$$\begin{aligned} M_\# &\le \liminf_{\lambda\to\infty} \frac{W(w_\lambda)}{J(w_\lambda)} \le \limsup_{\lambda\to\infty} \frac{W(w_\lambda)}{J(w_\lambda)} \\ &\le \limsup_{\lambda\to\infty} \frac{W(u_\lambda)}{\lambda} \le M_\#. \end{aligned}$$

Step 4. The sequence $\{u_\lambda\}$ is bounded in $H^4_{\mathrm{loc}}(\mathbb{R})$, and the sequence $\{w_\lambda\}$ is bounded in $H^2_{\mathrm{loc}}(\mathbb{R})$. This result depends crucially on the destiffening-restiffening character of F via the lemma below.

LEMMA 3.2. *Fix $K \in \mathbb{R}$. There exists $M > 0$ such that if $p \le K$ and $u \in L^\infty(\mathbb{R})$ solves* (1.5)*, then $\|u\|_{L^\infty(\mathbb{R})} \le M$.*

Note that the order of the quantifiers is important: the lemma states that if u is bounded, then it is bounded by a constant independent of u and p (subject to $p \le K$). We defer the proof of this lemma to the end of this section.

Since the functions u_λ satisfy (1.5) and $p_\lambda < 2$, the sequence $\{u_\lambda\}$ is bounded in $L^\infty(\mathbb{R})$. Standard elliptic estimates (e.g. [23, Theorem 11.1]) then give the boundedness of $\{u_\lambda\}$ in H^4 on compact sets. Since the cut-and-paste operation by which w_λ is constructed does not conserve H^4-regularity, the functions w_λ only enjoy the same regularity properties up to H^2-regularity.

As a consequence of the H^4-boundedness, u'_λ, u''_λ, and u'''_λ are all bounded in $L^\infty(\mathbb{R})$ independently of λ; additionally T_λ is bounded from below, since if $T_\lambda \to 0$, then by the bound on u'_λ, $\|u_\lambda\|_{L^\infty(0,T_\lambda)} = \|w_\lambda\|_{L^\infty(\mathbb{R})} \to 0$, so that we have $\liminf W(w_\lambda)/J(w_\lambda) \ge 2$. This contradicts (3.3).

Step 5. Convergence. Since w_λ is bounded in H^2_{loc} uniformly in λ, we can choose a sequence that converges weakly in $H^2_{\mathrm{loc}}(\mathbb{R})$ to a limit function w_∞.

1. If T_λ is bounded along this sequence, then—possibly after extracting a subsequence—T_λ and $J(w_\lambda)$ converge, and w_∞ is periodic with a finite period. The weak convergence implies that $W(w_\infty) \le \liminf W(w_\lambda)$, so that w_∞ is a solution of the minimization problem (3.1).

2. If T_λ is unbounded, then note that w_λ and u_λ have the same weak limit w_∞. We choose a subsequence such that p_λ, which is bounded between 0 and 2, converges. The weak convergence of u_λ in H^4_{loc} implies that w_λ satisfies (1.5) with limit load p_∞. This load lies necessarily between 0 and 2; this implies, as above, that solutions tending to zero oscillate around zero, contradicting the monotonicity of w_λ.

We conclude that case 2 does not occur.

If we pick $\delta > 0$ such that $(0,\delta)$ is included in $(0,T_\lambda/2)$ for all λ, and $\phi \in C_c^\infty((0,\delta))$, then

$$p_\lambda = \frac{W'(u_\lambda)\cdot\phi}{J'(u_\lambda)\cdot\phi} = \frac{W'(w_\lambda)\cdot\phi}{J'(w_\lambda)\cdot\phi} \to \frac{W'(w_\infty)\cdot\phi}{J'(w_\infty)\cdot\phi}$$

by the weak convergence of w_λ. By the remark made before the beginning of the proof, the fact that w_∞ minimizes the ratio W/J among all periodic functions implies that w_∞ also satisfies (1.5) with $p = M_\#$. Therefore $p_\lambda \to M_\#$.

Step 6. Conclusion. The functions u_λ and w_∞ solve the same differential equation (1.5) for loads p_λ and $M_\#$ that satisfy $p_\lambda \to M_\#$. We have $u_\lambda \rightharpoonup w_\infty$ in $H^2(0,\delta)$; using standard elliptic theory it follows that u_λ converges to w_∞ in $C^k(0,\delta)$ for all $k \in \mathbb{N}$. The classical result of continuous dependence on initial data then extends this to any compact set K. This concludes the proof of the theorem. ∎

We end this section with the proof of Lemma 3.2.

Proof. We first prove the lemma under the condition $|p| \le K$. Suppose that $p_n \in [-K,K]$ and u_n satisfy (1.5), with $\|u_n\|_{L^\infty(\mathbb{R})} \to \infty$. Set $\gamma_n = \|u_n\|^{-1}_{L^\infty(\mathbb{R})}$, so that $\gamma_n \to 0$, and define

$$v_n(x) = \gamma_n u_n(\gamma_n x).$$

Then

$$v_n'''' + p_n\gamma_n^2 v_n'' + \gamma_n^4 v_n - \gamma_n^2 v_n^3 + \alpha v_n^5 = 0.$$

Since v_n is uniformly bounded, classical elliptic estimates (e.g., [23]) imply that $v_n \rightharpoonup v_\infty$ in $H^4_{\mathrm{loc}}(\mathbb{R})$, after extraction of a subsequence. The limit v_∞ therefore satisfies the equation

$$v_\infty'''' + \alpha v_\infty^5 = 0 \qquad \text{on } \mathbb{R}, \tag{3.4}$$

which has no nonzero bounded solution (see, e.g., [19]). This contradicts the fact that $\|v_n\|_{L^\infty(\mathbb{R})} = 1$.

If we release the lower bound on p, and assume that $p_n \to -\infty$, then we define in addition

$$\delta_n = \max\{\gamma_n, |p_n|^{1/2}\gamma_n^2\}$$

and

$$v_n(x) = \gamma_n u_n(\delta_n x).$$

Since v'''' and pv'' are both positive operators if $p < 0$, the unboundedness of p is irrelevant for the elliptic estimates. The limit equation is

$$\overline{\gamma} v_\infty'''' - \overline{\delta} v_\infty'' + v_\infty^5 = 0 \qquad \text{on } \mathbb{R},$$

where $\overline{\gamma}, \overline{\delta} \in [0,1]$ and $\overline{\gamma} + \overline{\delta} \neq 0$. For none of the possible combinations of $\overline{\gamma}$ and $\overline{\delta}$ does this equation have a bounded nonzero solution. ∎

4. The periodic function $u_\#$. In the previous section we showed that there exists a solution $u_\#$ to the variational problem

$$M_\# = \inf\left\{\frac{W(u)}{J(u)} : u \in H^2_{\mathrm{loc}}(\mathbb{R}) \text{ periodic}\right\}$$

and that it is the limit, on compact sets, of solutions u_λ of problem (1.3). In this section we discuss a number of issues concerning this periodic function $u_\#$.

4.1. Critical buckling load. Going back to the model of an axially loaded strut, let us briefly examine the behavior under dead loading, rather than rigid loading; i.e., we fix the load p and seek an associated response. The appropriate energy for this loading situation is [24, p. 50]

$$\mathcal{L}(u) = W(u) - pJ(u), \tag{4.1}$$

which is often called the total potential or the Lagrangian. Note that equilibria of $\mathcal{L}$ again satisfy (1.4); both dead and rigid loading lead to the same equilibria, but the stability properties differ.

For small values of p, $\mathcal{L}$ is a positive definite function of u, and the trivial response, $u \equiv 0$, is the unique global minimizer. When p passes a threshold value there will be profiles with a negative Lagrangian, so that the zero response is no longer optimal, and can be improved upon by a nonzero deflection. Thus we can define a critical load p_c, such that

$$\begin{aligned} \inf_{u\in H^2(\mathbb{R})} \mathcal{L}(u) &= 0 \qquad \text{if } p < p_c, \\ \inf_{u\in H^2(\mathbb{R})} \mathcal{L}(u) &< 0 \qquad \text{if } p > p_c. \end{aligned}$$

Note that if $\inf \mathcal{L}(u) < 0$, then in fact $\inf \mathcal{L}(u) = -\infty$, by replication of an appropriate function u.

An alternative, but equivalent, way of representing the statements above is

$$p_c = \inf_{u\in H^2(\mathbb{R})} \frac{W(u)}{J(u)}.$$

Here the connection with the previous section becomes clear.

4.2. Symmetry of the minimizer. Variational problems very similar to that of $\inf \mathcal{L}$ arise in the study of polymeric materials under tension [11, 16]. It is interesting to note that the concept of a critical load (p_c), that has its origin in a mechanical viewpoint, is mirrored very closely by the ideas presented in [11], notably Theorem 6.1.

While the settings of [11, 16] are slightly different from the current one, some of the proofs carry over immediately. By adapting Lemmas 3.3 and 3.6 of [16] we find the following.

LEMMA 4.1 (see [16]).

1. *$u_\#$ is even about any critical point;*
2. *if $u_\#$ has a zero, then it is odd about this zero.*

As for the condition that $u_\#$ have zeros, this is easily proved as follows.

LEMMA 4.2. *$u_\#$ has a zero.*

Proof. Suppose that $u_\# > 0$ on $\mathbb{R}$. For $\mu > 1$, define $v_\mu = \max u_\# - \mu((\max u_\#) - u_\#)$. Then $\int v_\mu''^2 = \mu^2 \int u_\#''^2$, $\int v_\mu'^2 = \mu^2 \int u_\#'^2$, and $\int F(v_\mu) \le \int F(u_\#)$ provided

$v_\mu \geq 0$ (recall that $F'(u) \geq 0$ if $u \geq 0$). Therefore

$$\frac{W(v_\mu)}{J(v_\mu)} \leq \frac{\frac{\mu^2}{2}\int u_\#''^{\,2} + \int F(u_\#)}{\frac{\mu^2}{2}\int u_\#'^{\,2}} < \frac{W(u_\#)}{J(u_\#)},$$

which contradicts the minimality of $u_\#$. □

In summary, $u_\#$ is both odd and even.

5. Numerical computation of minimizers.

5.1. Procedure. The computation of global minimizers in a nonconvex setting suffers from the potential existence of a large number of local minimizers. The problem at hand—that of minimizing W for prescribed values of J—appears to be particularly demanding from this point of view, since the associated Euler–Lagrange equation (1.5) is expected to have a large number of homoclinic solutions. Champneys and Toland [4] showed the existence of a multitude of homoclinic orbits bifurcating from $p = -2$ for a related problem ($\alpha = 0$), which they numerically tracked into the $p > 0$ domain. These orbits are "multimodal," "repeated" versions of a primary orbit. In addition the existence of many "multibump" homoclinics has been shown, which consist of N copies of a given homoclinic, separated by large distances.

However, there is evidence that many of these homoclinic orbits are not constrained minimizers. There is a folk theorem, which received some backup in [22], that local stability under rigid loading is related to the change of J along an equilibrium path: if J decreases, then the solution is stable, and it is unstable otherwise. This would disqualify many equilibria off-hand. For the multibump homoclinics an additional argument suggests that they can never be stable (see again [22]). Based on this circumstantial evidence, we conjecture that the number of constrained local minimizers is in fact very limited. The numerical evidence of this section supports this conjecture, and we shall return to a further discussion of the issue in section 6.

We therefore adopt the following procedure to seek a global minimizer of problem (1.3) for given λ. Starting from quasi-random initial data (satisfying $J = \lambda$) we solve the constrained gradient flow problem

$$u_t = -u_{xxxx} - pu_{xx} - f(u), \qquad x \in \mathbb{R},\ t > 0, \tag{5.1}$$

$$J(u(\cdot, t)) = \lambda, \qquad t > 0. \tag{5.2}$$

Here $p = p(t)$ is a priori unknown, and is determined as part of the solution. This problem has a strictly decreasing Lyapunov function (the functional W), and converges rapidly to a stationary solution, which we assume to be a local minimizer. By repeating this process for a "large" number of different random initial data we collect a number of local minimizers. We select the solution with the lowest value of W as the global minimizer of W under the condition $J = \lambda$.

For the computation of solutions of the constrained dynamical system (5.1)–(5.2) we restrict the problem to a finite domain $(-L, L)$, with L suitably large, and impose the boundary conditions of a simply supported beam ($u(\pm L) = u_{xx}(\pm L) = 0$). An equivalent variational formulation follows by multiplying the equation by a test function v with $v(\pm L) = 0$ and integrating:

$$\int_{-L}^{L} u_t v\,dx + \int_{-L}^{L} u_{xx}v_{xx}\,dx - p\int_{-L}^{L} u_x v_x\,dx + \int_{-L}^{L} f(u)v\,dx = 0. \tag{5.3}$$

We now determine an approximation to $u(x,t)$ by using the finite-element method to give a semidiscretization of (5.3) [25]. To do this we approximate $u(x,t)$ by the function $U_h(x,t) = \sum U_i(t)\phi_i(x) + \sum U_{xi}(t)\psi_i(x)$. Here ϕ_i and ψ_i are piecewise cubic functions defined on a uniform mesh of spacing $h := 2L/N$ so that

$$\phi_i(-L+jh) = \psi_i'(-L+jh) = \delta_{ij} \qquad \text{and} \qquad \psi_i(-L+jh) = \phi'(-L+jh) = 0 \quad \text{for } i,j = 0,\dots,N.$$

The space S_h is the span of the functions ϕ_i $(i = 1,\dots,N-1)$ and ψ_i $(i = 0,\dots,N)$ (such that the imposed boundary condition $u = 0$ is incorporated into the solution space). We set $U \in \mathbb{R}^{2N}$ equal to $U = U(t) = (U_1,\dots,U_{N-1},U_{x0},\dots,U_{xN})$. Now we require that U_h should satisfy (5.3) for all functions $V \in S_h$. Setting $V = \phi_i$ or $V = \psi_i$ leads to the following system of ODEs for U and P:

$$AU_t + BU - PCU + D = 0, \tag{5.4}$$

where the $2N \times 2N$ matrices A, B, and C are given by

$$A_{ij} = \int \phi_i\phi_j, \qquad 1 \le i,j \le N-1,$$
$$B_{ij} = \int \phi_i''\phi_j'', \qquad 1 \le i,j \le N-1,$$
$$C_{ij} = \int \phi_i'\phi_j', \qquad 1 \le i,j \le N-1,$$

with similar entries for other ranges of i and j. The components D_i of the zero-order term D in (5.4) are numerical approximations, using Simpson's rule, of the integral

$$\int f(U_h)\phi_i, \qquad 1 \le i \le N-1,$$
$$\int f(U_h)\psi_{i-N}, \qquad N \le i \le 2N.$$

The in-plane load $p(t)$ is determined as part of the solution and the necessary and sufficient condition comes from the integral constraint (5.2), which reads in discretized form

$$\frac{1}{2}U^TCU = \lambda. \tag{5.5}$$

The system (5.4)–(5.5) is then an index-2 differential-algebraic equation. Differentiating (5.5) with respect to time we find

$$U^TCU_t = 0. \tag{5.6}$$

We solved (5.4) and (5.6) using `DDASSL`, a backward-difference form differential-algebraic equation solver [20]. We choose to replace the constraint (5.5) by (5.6) since the latter provides a DAE system of index one, which `DDASSL` is designed to handle. It is verified after calculation that the deviation from (5.5) due to accumulation of numerical error is acceptably small (relative error less than 0.01).

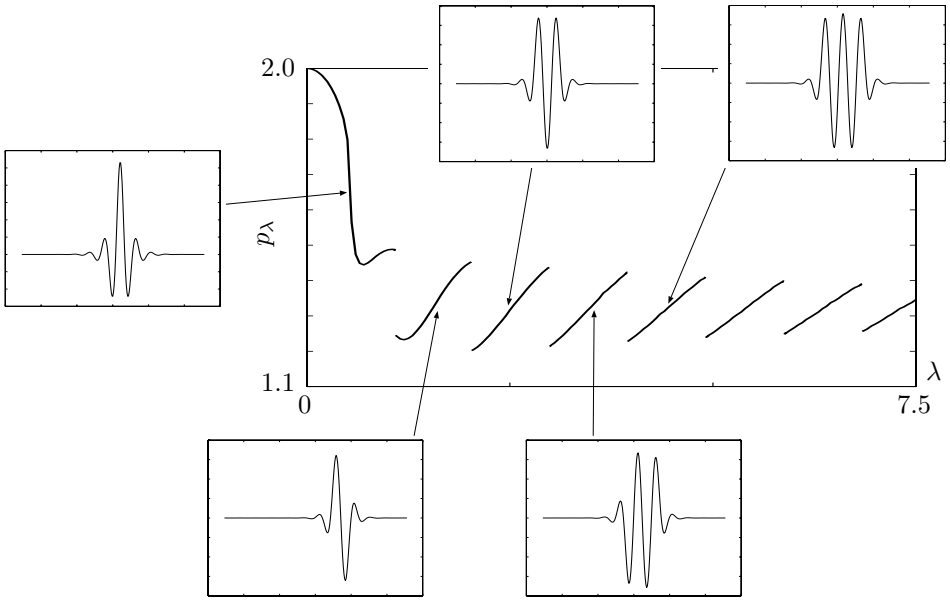


FIG. 5.1. *Results of the numerical minimization ($\alpha = 0.3$).*

5.2. Results. Figure 5.1 shows a plot of the load p_λ as a function of λ. The initial data sample size is 25.

A number of features of this graph merit special mention.

1. The graph decomposes into a collection of continuous curves. The apparent discontinuities in this figure are actual discontinuities; the change in λ causes local minima to move relative to each other, and at these discontinuities the global minimum jumps from one local minimum to another. Also, it appears that the continuous curves are projections of continua of solutions in state space (note that comparison is not trivial because of the interference of the discretization; also, we do not want to impose any symmetry).

2. Theorem 3.1 states that for any sequence $\lambda_n \to \infty$, $p_{\lambda_n} \to M_\#$. In Figure 5.1 we recognize this convergence in the decrease of the vertical extent of the graph as λ increases.

3. On the continuous parts of the curve, the solution has either odd or even symmetry. At the jumps the solution switches from one to the other.

4. The load is not a continuous function of λ; but all jumps are downward. Compare this to Lemma 2.6.

In the next section we give an interpretation of the form of Figure 5.1.

6. Correspondence with the bifurcation diagram. In this section we briefly change our perspective: instead of problem (1.3) we consider the ODE (1.5),

$$u'''' + pu'' + f(u) = 0 \qquad \text{on } \mathbb{R}, \tag{1.5}$$

where p is a prescribed parameter. A solution of (1.3) also solves (1.5), but the opposite is not true. As we mentioned in the previous section, there are many solutions of (1.5) that are strongly suspected of not even being local constrained minimizers.

Figure 6.1 shows a bifurcation plot of (1.5). At $p = 2$, at zero J, a Hamiltonian–Hopf bifurcation creates four homoclinic orbits. Two of these are even, and each the

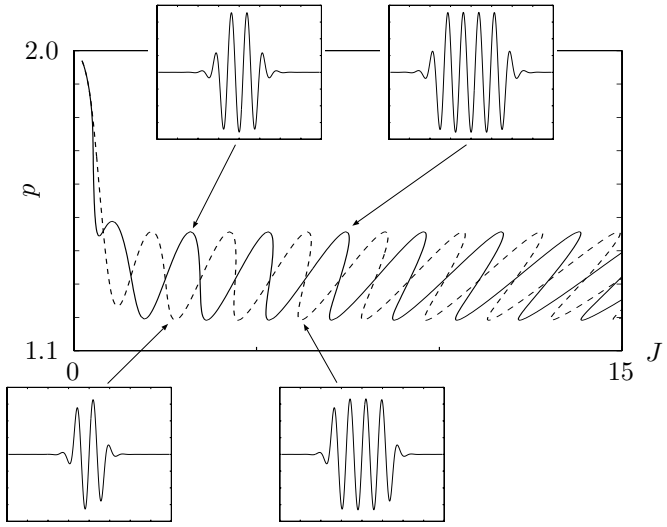


FIG. 6.1. *Bifurcation diagram for* (1.5) *showing curves of even (continuous line) and odd (dashed line) solutions bifurcating from* $p = 2$. *Here* $\alpha = 0.3$.

opposite of the other ($u_2 = -u_1$); the other two are odd, and again each other's opposite. In Figure 6.1 we identify the two even and the two odd solutions and thus draw two curves in total.

The initial part of the figure, near $p = 2$, is typical for a destiffening nonlinearity. The oscillating behavior for larger values of J, however, is related to the competing destiffening and restiffening qualities. It is shown in [19] how the restiffening nature (more specifically, the fact that $F(u) > F(0)$ for $u \neq 0$) implies that along the curve p must be bounded from below. Woods [26] and Woods and Champneys [27] show that the snaking behavior can be explained as the result of a collision of the unstable manifold of zero with the stable manifold of a family of periodic orbits parametrized by p. When $p = M_{\#}$, this periodic orbit is exactly the function $u_{\#}$ of Theorem 3.1.

When we combine this figure and Figure 5.1 into one diagram (Figure 6.2) there is a strong suggestion that all minimizers lie on the bifurcation curve. If we elevate this numerical suggestion to the status of hypothesis, that is, if we suppose that all minimizers of problem (1.3) lie on this bifurcation diagram, then the jumps from one curve to the other result from a simple energy argument. In a graph of load against deflection, strain energy is represented by area under the graph. More precisely, if we have a continuum of solutions v_s of (1.4), parametrized by s, with associated load p_s, then

$$\begin{aligned} W(v_{s_2}) - W(v_{s_1}) &= \int_{s_1}^{s_2} W'(v_s) \cdot \frac{dv_s}{ds}\, ds \\ &= \int_{s_1}^{s_2} p_s J'(v_s) \cdot \frac{dv_s}{ds}\, ds \\ &= \int_{s_1}^{s_2} p_s\, dJ(v_s), \end{aligned}$$

with a slight abuse of notation in the last integral.

To explain the jumps, let us assume, to start with, that for some interval of

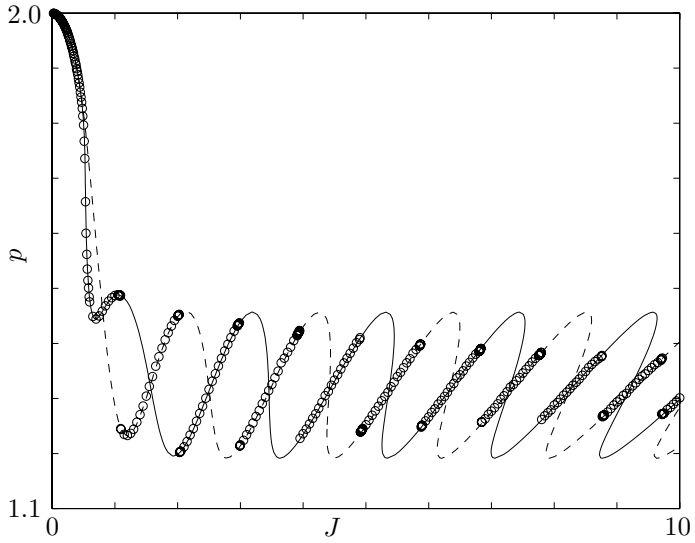


FIG. 6.2. *Combination of Figures* 5.1 *and* 6.1.

values of λ all minimizers lie on a given continuum of solutions v_s. This is shown schematically in Figure 6.3. At the critical value λ_c the two areas E_1 and E_2 are equal, implying that the strain energies at **A** and **B** are equal. As λ passes through the critical value, the minimum in the strain energy jumps from the top to the bottom curve.

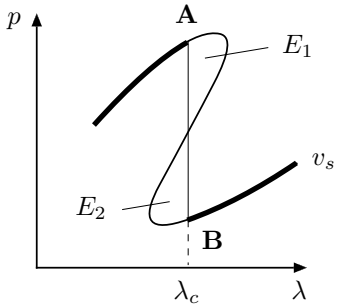


FIG. 6.3. *The thick line indicates the minimizer under constrained* λ.

In the case of the problem as stated in (1.3), the numerical results clearly indicate that both the branches of solutions in Figure 6.1 contain minimizers. We therefore need to take both curves into account when searching for jumps. As an example, Figure 6.4 shows a blow-up of the first jump in Figure 5.1, where the minimum passes from the even to the odd branch. Again the jump corresponds to an equal-area condition. The other jumps arise in the same manner.

In summary, if we make the assumption that all global minimizers of (1.3) lie on the bifurcation diagram of Figure 6.1, then the form of Figure 5.1 follows readily from energy comparison.

The assumption that all global minimizers lie on the bifurcation diagram is a strong one. As of yet there is no conclusive argument why this might be the case. For some specific classes of solutions of (1.5) it has been shown that they are or are not locally minimal (see above) but these results depend in a critical manner on the

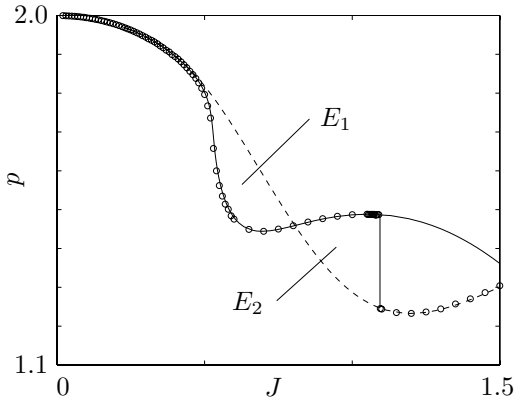


FIG. 6.4.

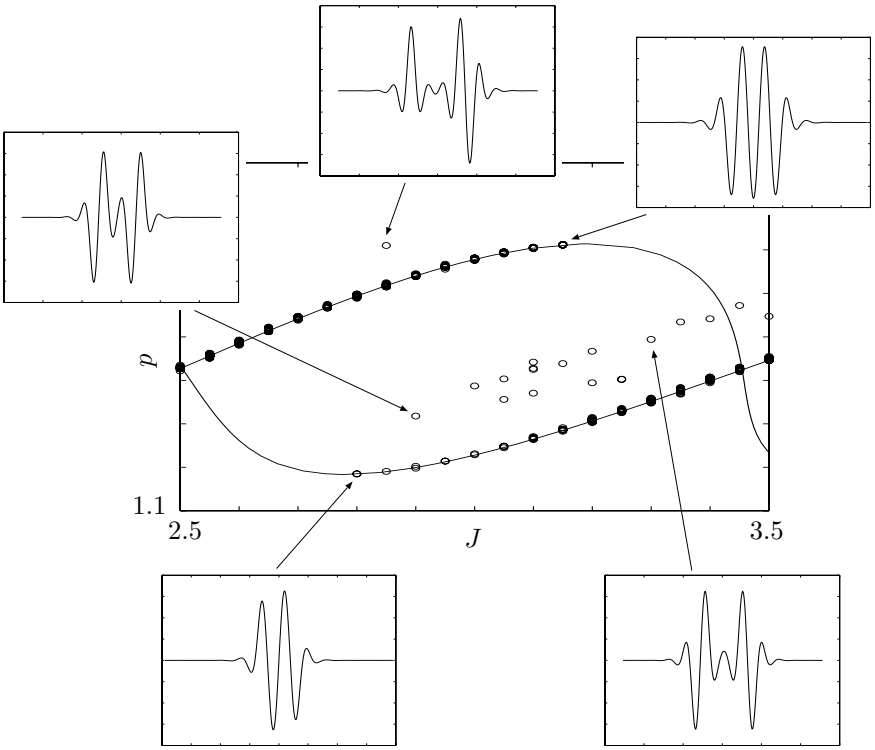


FIG. 6.5. *Every circle represents a "local minimizer" that was found numerically (see text).*

structure of the solutions involved. A complete classification of all solutions of (1.5) is still a distant goal, and therefore doing an exhaustive search is not an option.

To complicate matters, the numerical results suggest that local optimality does not guarantee membership of the bifurcation diagram in Figure 6.1. As mentioned before, the algorithm used for finding the global minimizer runs a constrained gradient flow algorithm starting from random initial data; the function that the algorithm stabilizes at for large time is assumed to represent a local minimum. This procedure is repeated a number of times, and the local minimum with least strain energy is

tagged as the global minimum. This is the solution that appears in Figures 5.1 and 6.2.

In Figure 6.5 we show an excerpt of Figure 6.2, but this time we plot not only the global minimum but all of the local minima that were found along the way. In addition to the solutions that we would expect, those that lie on the two bifurcation curves, other solutions appear with a different structure. Of course, "local optimality" has been established in a crude manner, so this could well be a numerical artifact. The question of the relationship between the minimization problem (1.3) and Figure 6.1 remains an interesting one, however, that merits being studied in more detail.

7. The nonlinearity F. In this paper we concentrate entirely on functions F of the form (1.6). Of course the class of functions for which one can derive the same results is much larger, and in this section we give some indication as to which properties of F enter into play. In addition, the existence result (Theorem 2.2) and the convergence result (Theorem 3.1) differ in their requirements, and we shall also comment on this issue.

The term destiffening was defined in the introduction as a decrease in the marginal stiffness $F''(u)$ as u moves away from zero. The actual property used in the proofs, however, is the combination "$F'''(0) = 0$ and $F''''(0) < 0$" (in Lemma 2.3, part 1). The function (1.6) satisfies both of these formulations of the destiffening character, but the function $F(u) = u^2/2 - u^6/6 + \alpha u^8/8$, for instance, satisfies only the first of the two. As remarked on page 1149, the proof of lemma 2.3 does not apply to this latter function, but numerical results suggest that the assertion of the Lemma ($W_\lambda < 2\lambda$) holds nonetheless. At this stage we must conclude that there is a grey area between these two formulations of "destiffening."

If we tolerate this lack of accuracy for the moment, we can assert that the destiffening nature is crucial for the existence proof, via the same property $W_\lambda < 2\lambda$ and the estimate (2.11). However, destiffening alone is not sufficient to guarantee existence for all $\lambda > 0$. If F takes negative values (assuming $F(0) = 0$), say $F(\bar{u}) < 0$, then for sufficiently large values of λ we can create admissible profiles with large *negative* strain energy. As an example, consider

$$u_k(x) = \bar{u}\,\eta(|x| - k),$$

where η is a smooth cut-off function such that $\eta \equiv 1$ on $(-\infty, -1]$ and $\eta \equiv 0$ on $[1, \infty)$. If $k > 1$, then $J(u_k)$ is independent of k, but $W(u_k)$ takes arbitrarily large negative values as $k \to \infty$. Since we therefore have $\inf_{\mathcal{C}_\lambda} W = -\infty$, the existence question is absurd. In order to avoid this degeneracy, we need to assume $F(u) \geq 0$ (the possibility $F(\bar{u}) = 0$, $\bar{u} \neq 0$ leads to noncompactness of minimizing sequences; however, such sequences can be adapted to regain compactness, so that the existence of a minimizer is not compromised).

To summarize, the main characteristics of F that lead to existence are the destiffening nature and this positivity property. The function (1.6) meets these constraints if and only if $\alpha \geq 3/16$.

Turning to the convergence of minimizers as the end-shortening λ tends to infinity (Theorem 3.1), simple positivity of F is not sufficient. One can construct counter-examples where $F(u)$ is small, but positive, for large $|u|$; minimizers for such nonlinearities are unbounded in the L^∞-norm and therefore do not converge. Some form of stiffening for larger u is necessary to prevent this runaway. As before, no sharp condition is known, but Lemma 3.2, which provides the all-important L^∞ bound, can

be proved for all F with

$$F'(s) \sim s^{q-1} \qquad \text{as } |s| \to \infty, \qquad \text{with } q > 2,$$

without any change in the proof. For such functions F the statement of Theorem 3.1 should hold unchanged.

In addition, for the convergence result we assume that $\alpha \geq 1/4$, so that $p \geq 0$ (see (3.2)), and solutions necessarily oscillate at infinity. This property is used twice in the proof of Theorem 3.1. We conjecture that $\alpha \geq 1/4$ is unnecessarily restrictive, however, and that $\alpha \geq 3/16$ should suffice for both existence of minimizers and the convergence for large λ.

While dwelling on the subject of the nonlinearity F, we might also comment on the requirement of restiffening itself, i.e., the fact that we assume a relatively complex structure in the response of the elastic foundation. It is true that the combination of initial destiffening and subsequent stiffening appears artificial. However, there is good reason to assume that both the destiffening and the subsequent stiffening characters are present in actual examples of elastic struts on foundations—not in the foundation response, but in other elements in the model. For instance, in linearizing the higher-order terms in the equation—that is, by replacing $\mathcal{W}$ and $\mathcal{J}$ by W and J—a destiffening property that is present in the original formulation has been discarded. Mühlhaus [17] gives a heuristic argument for this fact, and it can be verified by doing a small-amplitude development of the appropriate nonlinear terms.

Similarly, a foundation that does not have the local response of the Winkler foundation that we consider here, but "feels" the proximity of the layer at adjoining sites, has a strongly stiffening character for large deformations. This is illustrated by Figure 7.1, where the material indicated by the hashing, being squashed by the bends in the strut, will exert a large force on the strut in the opposite direction. This is an inherently nonlocal effect that cannot be captured with a Winkler foundation. In summary, the various simplifying assumptions that we have made during the modelling process have removed the destiffening and subsequent stiffening characteristics from the formulation, forcing us to reintroduce them via the foundation response.

With these arguments in mind we chose to consider a mathematical model that has the nature, if not the exact form, of the mechanical problem. We hope that the ideas of this paper will be amenable to future extension.

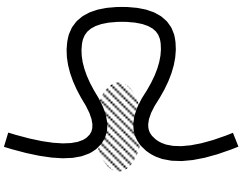


FIG. 7.1. *Squashed material exerts a nonlocal force on the strut.*

Appendix. Proof of Lemma 2.5. It is relatively simple to prove that for any minimizer u the load necessarily satisfies $p \leq 2$. If u minimizes W at constant J, with associated load p, then u is a stationary point of the functional

$$\mathcal{L}(u) = W(u) - pJ(u).$$

Since the constraint is one-dimensional, the second derivative of $\mathcal{L}$ at u cannot have more than one unstable eigenmode.

On the other hand, suppose that $p > 2$ and let $\phi \in C_c^\infty(\mathbb{R})$ satisfy

$$\int \left(\phi''^2 - p\phi'^2 + \phi^2\right) < 0.$$

If $|x|$ is sufficiently large, then $F''(u(x)) \leq 1$ and therefore, setting $\psi_K(x) = \phi(x-K)$,

$$\mathcal{L}''(u)\cdot\phi_K\cdot\phi_K = \int \left({\phi_K''}^2 - p{\phi_K'}^2 + F''(u){\phi_K}^2\right) < 0,$$

both for large and for small K. It follows that $\mathcal{L}$ has at least two unstable directions, and this contradicts the assumption that u is a minimum.

When we write (1.5), for $p = 2$, as a four-dimensional dynamical (Hamiltonian) system, then the linear part of this system is given by a matrix which is not diagonizable. Using normal form theory, for every $k \in \mathbb{N}$ we can transform the system to a system given by the Hamiltonian

$$H = \frac{1}{2}|p|^2 + \langle p, Jq\rangle + P(|q|^2, \langle p, Jq\rangle) + O(|p|^{2k}, |q|^{2k}) \tag{A.1}$$

(see, e.g., [15, Chapter VII]). Here P is a polynomial in its two arguments, whose lowest-order terms are quadratic. For our purposes the only important term in $P(u,v)$ is au^2, or equivalently $a|q|^4$. The calculations done by Woods [26] show that $a > 0$.

By hypothesis, the orbit represented by $(p(t), q(t))$ converges to the origin as $t \to \infty$. Since the system is linear in the limit of small amplitude, it follows that (p,q) must converge to solutions of the linear problem. More accurately, if we choose $t_n \to \infty$, and rescale by setting

$$(p_n, q_n)(t) = \frac{1}{|(p,q)(t_n)|}(p,q)(t - t_n),$$

then the functions (p_n, q_n) converge on compact subsets to bounded solutions of the linear problem. Since all such solutions satisfy $p \equiv 0$, it follows that $p = o(|q|)$ as $t \to \infty$.

We next transform (p,q) to polar coordinates (r, R, θ, Θ), given by

$$\begin{aligned} q_1 &= r\cos\theta, & q_2 &= r\sin\theta,\\ p_1 &= R\cos\theta - \left(\frac{\Theta}{r}\right)\sin\theta, & p_2 &= R\sin\theta + \left(\frac{\Theta}{r}\right)\cos\theta. \end{aligned}$$

The Hamiltonian then takes the form

$$H(r,R,\theta,\Theta) = \frac{1}{2}\left(R^2 + \left(\frac{\Theta}{r}\right)^2\right) - \Theta + P(r^2,\Theta) + O\left(\left|R^2 + \left(\frac{\Theta}{r}\right)^2\right|^k + r^{2k}\right).$$

The result $p = o(|q|)$ translates to $R/r, \Theta/r^2 \to 0$, which implies, together with $H = 0$, that

$$\Theta \sim \frac{1}{2}R^2 + ar^4.$$

We can then calculate an estimate of the rate of decay of Θ:

$$\dot{\Theta} = -\frac{\partial H}{\partial \theta} = O(R^{2k} + r^{2k}) = O(\Theta^{k/2}).$$

It follows that for $k \geq 4$ the rate of decay of Θ is too small to be compatible with the condition that $u \in H^2(\mathbb{R})$, or

$$\int^{\infty} \left(|p|^2 + |q|^2\right) < \infty,$$

since

$$|p|^2 + |q|^2 \geq R^2 + r^2 \geq c\Theta$$

for some $c > 0$, in the limit $t \to \infty$.

Acknowledgments. Many of the ideas presented in this paper arose in discussions at the Centre for Nonlinear Mechanics, Bath, and the author is grateful to Chris Budd and Giles Hunt for providing such a stimulating environment. In addition, the author is indebted to Boris Buffoni for the idea of Lemma 2.5.

REFERENCES

[1] M. A. Biot, *Mechanics of Incremental Deformations*, Wiley, New York, 1965.
[2] A. R. Champneys, *Homoclinic orbits in reversible systems and their applications in mechanics, fluids and optics*, Phys. D, 112 (1998), pp. 158–186.
[3] A. R. Champneys, *Homoclinic orbits in reversible systems* II: *Multi-bumps and saddle-centres*, CWI Quarterly, 12 (1999), pp. 185–212.
[4] A. R. Champneys and J. F. Toland, *Bifurcation of a plethora of multi-modal homoclinic orbits for autonomous Hamiltonian systems*, Nonlinearity, 6 (1993), pp. 665–772.
[5] P. R. Cobbold, *Fold progatation in single embedded layers*, Tectonophysics, 27 (1975), pp. 333–351.
[6] P. R. Cobbold, *Finite-element analysis of fold propagation—a problematic application?*, Tectonophysics, 38 (1977), pp. 339–353.
[7] G. W. Hunt and P. R. Everall, *Arnold tongues and mode-jumping in the supercritical post-buckling of an archetypal elastic structure*, R. Soc. Lond. Proc. A, (1998), pp. 125–140.
[8] G. W. Hunt, G. J. Lord, and A. R. Champneys, *Homoclinic and heteroclinic orbits underlying the post-buckling of axially-compressed cylindrical shells*, Comp. Methods Appl. Mech. Engrg., 170 (1999), pp. 239–251.
[9] G. W. Hunt, M. A. Peletier, A. R. Champneys, P. D. Woods, M. A. Wadee, C. J. Budd, and G. L. Lord, *Cellular buckling in long structures*, Nonlinear Dynam., 21 (2000), pp. 3–29.
[10] G. W. Hunt, M. A. Peletier, and M. A. Wadee, *The maxwell stability criterion in pseudo-energy models of kink banding*, J. Structural Geology, 22 (2000), pp. 667–679.
[11] A. Leizarowitz and V. J. Mizel, *One dimensional infinite-horizon variational problems arising in continuum mechanics*, Arch. Rational Mech. Anal., 106 (1989), pp. 161–194.
[12] P.-L. Lions, *The concentration-compactness principle in the calculus of variations. The locally compact case* I. Ann. Inst. Henri Poincaré Anal. Non Linéaire, 1 (1984), pp. 109–145.
[13] P.-L. Lions, *The concentration-compactness principle in the calculus of variations. The locally compact case* II. Ann. Inst. Henri Poincaré Anal. Non Linéaire, 1 (1984), pp. 223–283.
[14] G. J. Lord, A. R. Champneys, and G. W. Hunt, *Computation of localized post buckling in long axially-compressed cylindrical shells*, Philos. Trans. Roy. Soc. London Ser. A, 355, 1997, pp. 2137–2150.
[15] K. R. Meyer and G. R. Hall, *Introduction to Hamiltonian Dynamical Systems and the N-body Problem*, Springer-Verlag, New York, 1992.
[16] V. J. Mizel, L. A. Peletier, and W. C. Troy, *Periodic phases in second-order materials*, Arch. Rational Mech. Anal., 145 (1998), pp. 343–382.
[17] H. B. Mühlhaus, *Evolution of elastic folds in plane strain*, in Modern Approaches to Plasticity, D. Kolymbas, ed., Elsevier Science Publishers B.V., New York, 1993, pp. 737–765.

[18] L. A. Peletier and W. C. Troy, *Spatial patterns in higher order phase transitions*, CWI Quarterly, 9 (1996), pp. 121–130.
[19] M. A. Peletier, *Non-existence and uniqueness for the extended Fisher-Kolmogorov equation*, Nonlinearity, 12 (1999), pp. 1555–1570.
[20] L. R. Petzold, *A Description of* `DDASSL`*: A Differential/Algebraic System Solver*, Tech. Rep. SAND82-8637, Sandia National Laboratories, Livermore, CA, 1982.
[21] N. J. Price and J. W. Cosgrove, *Analysis of Geological Structures*, Cambridge University Press, London, 1990.
[22] B. Sandstede, *Instability of localized buckling modes in a one-dimensional strut model*, Philos. Trans. Roy. Soc. London Ser. A, 355 (1997), pp. 2083–2097.
[23] M. E. Taylor, *Partial Differential Equations* I*, Basic Theory*, Springer-Verlag, New York, 1996.
[24] J. M. T. Thompson and G. W. Hunt, *A General Theory of Elastic Stability*, Wiley, London, 1973.
[25] R. Wait and A. R. Mitchell, *Finite Element Analysis and Applications*, Wiley, New York, 1985.
[26] P. D. Woods, *Localisation in Fourth-Order Ordinary Differential Equations*, Ph.D. thesis, Department of Engineering Mathematics, University of Bristol, UK, 1999.
[27] P. D. Woods and A. R. Champneys, *Heteroclinic tangles and homoclinic snaking in the unfolding of a degenerate reversible Hamiltonian Hopf bifurcation*, Phys. D, 129 (1999), pp. 147–170.
[28] N. Yamaki, *Elastic Stability of Circular Cylindrical Shells*, North Holland Ser. Appl. Math. Mech., North-Holland, Amsterdam, 1984.

SIAM J. MATH. ANAL.
Vol. 32, No. 5, pp. 1169–1170

ERRATUM TO "WEIGHTED ZERO DISTRIBUTION FOR POLYNOMIALS ORTHOGONAL ON AN INFINITE INTERVAL"*

WALTER VAN ASSCHE†

Abstract. In [W. Van Assche, *SIAM J. Math. Anal.*, 16 (1985), pp. 1317–1334] the second term in the asymptotic expansion of Laguerre polynomials (Perron's formula) was computed and used to obtain some information about the rate of convergence of weighted zero distributions for Laguerre and Hermite polynomials. This second term in Theorem 2.3 and in the appendix of [W. Van Assche, *SIAM J. Math. Anal.*, 16 (1985), pp. 1317–1334] is in error. In this note we give the correct expression for this second term and indicate how this affects the results on the rate of convergence for the weighted zero distributions.

Key words. Laguerre polynomials, Perron's formula, asymptotic expansion

AMS subject classification. 33C25

PII. S0036141099359871

1. Perron's formula for Laguerre polynomials. Thomas Müller [1, pp. 548–549] recently expressed doubt about the correctness of the second term in the asymptotic expansion for Laguerre polynomials (Perron's formula) as given in Theorem 2.3 in [3]. Perron's formula for Laguerre polynomials is for $\alpha > -1$

$$L_n^{(\alpha)}(z) = \frac{1}{2\sqrt{\pi}} e^{z/2} (-z)^{-(2\alpha+1)/4} n^{(2\alpha-1)/4} \exp(2\sqrt{-nz}) \times \left(\sum_{j=0}^{p-1} C_j(\alpha; z) n^{-j/2} + \mathcal{O}(n^{-p/2}) \right), \tag{1.1}$$

where the bound for the remainder holds uniformly on every compact subset of $\mathbb{C} \setminus [0, \infty)$, $(-z)^{-(2\alpha+1)/4}$ and $\sqrt{-z}$ must be taken to be real and positive if $z < 0$. It is known that $C_0(\alpha; z) = 1$. The second term $C_1(\alpha; z)$ was worked out in the appendix of [3], but unfortunately an error was made in working out the expansion of the term $A_0(z)$ on page 1333. The correct expansion for $A_0(z)$ should read

$$A_0(z) = \frac{1}{2} (-z)^{-1/4} (-1)^{\alpha/2} n^{(2\alpha-1)/4} \exp(2\sqrt{-nz}) e^{-z/2} \times \left\{ 1 + \frac{\sqrt{-z}}{2\sqrt{n}} \left(\alpha + 1 - \frac{z}{6} \right) + \mathcal{O}\left(\frac{1}{n} \right) \right\}. \tag{1.2}$$

The resulting expression for $C_1(\alpha; z)$ then becomes

$$C_1(\alpha; z) = \frac{1}{4\sqrt{-z}} \left(\frac{1}{4} - \alpha^2 - 2(\alpha + 1)z + \frac{z^2}{3} \right). \tag{1.3}$$

*Received by the editors August 3, 1999; accepted for publication July 19, 2000; published electronically February 21, 2001.
http://www.siam.org/journals/sima/32-5/35987.html
†Katholieke Universiteit Leuven, Department of Mathematics, Celestijnenlaan 200 B, B-3001 Leuven, Belgium (walter@wis.kuleuven.ac.be). The author is Research Director of the Belgian National Fund for Scientific Research.

2. Weighted zero distributions for Laguerre and Hermite polynomials. The explicit expression of the term $C_1(\alpha;z)$ was only used in section 4 of [3], in particular in Theorems 4.1, 4.3, and 4.4. The proofs of these three theorems remain valid and Theorems 4.1 and 4.3 are correct as formulated. The formulation of Theorem 4.4 needs to be changed to the following.

THEOREM 4.4.

(i) *Let $Z_n^+(x)$ and $Z^+(x)$ be as in the case of generalized Laguerre polynomials and put $R_n^+(x)=\sqrt{n}\{Z_n^+(x)-Z^+(x)\}$. Let $f(x)$ be such that $f((1+y)/(1-y))$ is analytic in some open set containing $[-1,1]$; then as $n\to\infty$*

$$\sum_{j=1}^{n}\frac{f(x_{j,n}^+)}{1+x_{j,n}^+}-\frac{\sqrt{n}}{\pi}\int_0^\infty\frac{f(x)}{\sqrt{x}(1+x)}\,dx=\int_0^\infty f(x)\,dR_n^+(x)$$
$$\to-\frac{2\alpha+1}{4}f(0)-\frac{1}{2}f(\infty).$$

(ii) *Let $Z_n(x)$ and $Z(x)$ be as in the case of generalized Hermite polynomials and put $R_n^+(x)=\sqrt{2n}\{Z_n(x)-Z(x)\}$. Let $f(x)$ be a function such that $f(\pm((1+y)/(1-y))^{1/2})$ is analytic in some open set containing $[-1,1]$; then as $n\to\infty$*

$$\sum_{j=1}^{n}\frac{f(x_{j,n})}{1+x_{j,n}^2}-\frac{\sqrt{2n}}{\pi}\int_{-\infty}^\infty\frac{f(x)}{1+x^2}\,dx=\int_{-\infty}^\infty f(x)\,dR_n(x)$$
$$\to-\alpha f(0)-\frac{1}{2}f(\infty)-\frac{1}{2}f(-\infty).$$

Acknowledgments. The author thanks Thomas Müller (Queen Mary and Westfield College, London) for pointing out the error and Frank Olver (University of Maryland) for providing two alternative ways to compute the required term $C_1(\alpha,z)$ using the techniques from his book [2].

REFERENCES

[1] T. MÜLLER, *Finite group actions and asymptotic expansion of $e^{P(z)}$*, Combinatorica, 17 (1997), pp. 523–554.

[2] F. W. J. OLVER, *Asymptotics and Special Functions*, A. K. Peters, Wellesley, MA, 1997; originally published by Academic Press, New York, 1974.

[3] W. VAN ASSCHE, *Weighted zero distribution for polynomials orthogonal on an infinite interval*, SIAM J. Math. Anal., 16 (1985), pp. 1317–1334.

SIAM J. MATH. ANAL.
Vol. 32, No. 6, pp. 1171–1197

VARIATIONAL APPROXIMATION OF A SECOND ORDER FREE DISCONTINUITY PROBLEM IN COMPUTER VISION*

LUIGI AMBROSIO†, LORIS FAINA‡, AND RICCARDO MARCH§

Abstract. We consider a functional, proposed by Blake and Zisserman for computer vision problems, which depends on free discontinuities, free gradient discontinuities, and second order derivatives. We show how this functional can be approximated by elliptic functionals defined on Sobolev spaces. The approximation takes place in a variational sense, the De Giorgi Γ-convergence, and extends to this second order model an approximation of the Mumford–Shah functional obtained by Ambrosio and Tortorelli. For the purpose of illustration an algorithm based on the Γ-convergent approximation is applied to the problem of computing depth from stereo images and some numerical examples are presented.

Key words. theory and algorithms for computer vision, variational problems, Γ-convergence, functions of bounded variation

AMS subject classifications. 46E30, 49J45

PII. S0036141000368326

1. Introduction.

In recent years, variational principles with a free discontinuity set have been introduced to solve reconstruction problems in computer vision theory (see, for instance, [4, 26, 31]). The variational approach to the image segmentation problem proposed by Mumford and Shah [27] consists of minimizing the functional

$$E(u,K)=\int_{\Omega\setminus K}\left(|\nabla u|^2+\mu|u-g|^2\right)\,dx+\alpha\mathcal{H}^{n-1}(K\cap\Omega)\ , \tag{1.1}$$

where $\Omega\subset\mathbf{R}^n$ is a bounded open set, $\mathcal{H}^{n-1}$ is the Hausdorff $(n-1)$-dimensional measure, $g\in L^\infty(\Omega)$, and α, $\mu>0$ are fixed positive parameters. The functional has to be minimized over all closed sets $K\subset\overline{\Omega}$ and all $u\in C^1(\Omega\setminus K)$. In the case $n=2$ the function g represents the image to be segmented. By minimizing the functional one tries to detect the discontinuities of g due to the edges of the objects in the image, and to cancel the discontinuities due to noise and small irregularities. The set K contains the jump points of u and represents the edges of the objects. The functional penalizes large sets K, and outside K the function u is required to be close to g and C^1.

The Mumford and Shah variational principle can be extended to several reconstruction problems of computer vision [25]: stereo reconstruction [32], computation of optical flow [28], shape from shading [33]. Variational problems involving functionals of this form are usually called free discontinuity problems, after a terminology introduced by De Giorgi [18].

*Received by the editors February 22, 2000; accepted for publication (in revised form) October 3, 2000; published electronically February 28, 2001.
http://www.siam.org/journals/sima/32-6/36832.html

†Scuola Normale Superiore, Piazza dei Cavalieri 7, 56126 Pisa, Italy (luigi@ambrosio.sns.it). The work of the first author was partially supported by Progetto Nazionale MURST "Equazioni Differenziali e Calcolo delle Variazioni" 9701226040.

‡Dipartimento di Matematica, Via L. Vanvitelli 1, 06123 Perugia, Italy (faina@unipg.it).

§Istituto per le Applicazioni del Calcolo, CNR, Viale del Policlinico 137, 00161 Roma, Italy (march@iac.rm.cnr.it). The work of the third author was partially supported by Progetto MURST 5% "Rete Multimediale, LdR: Applicazione ai Beni Culturali."

The Mumford and Shah model has some drawbacks: it is unable to reconstruct crease discontinuities and yields the over-segmentation of steep gradients (the so-called ramp effect). To overcome these defects of the first order model, Blake and Zisserman [9] introduced a second order functional which can be written in the form

$$F(u,K_0,K_1)=\int_{\Omega\setminus(K_0\cup K_1)}\left(|\nabla^2 u|^2+\Phi(x,u)\right)\,dx + \alpha\mathcal{H}^{n-1}(K_0\cap\Omega)+\beta\mathcal{H}^{n-1}((K_1\setminus K_0)\cap\Omega)\ , \tag{1.2}$$

with $\alpha,\,\beta>0$ positive parameters. The functional has to be minimized over the unknown sets K_0, K_1, with $K_0\cup K_1$ closed and $u\in C^2(\Omega\setminus(K_0\cup K_1))$ approximately continuous on $\Omega\setminus K_0$. If $\Phi(x,u)=\mu|u-g|^2$ and $n=2$, the functional (1.2) is just that one introduced in [9] (the thin plate surface under tension). In the second order model, K_0 represents the set of jump points for u, and $K_1\setminus K_0$ is the set of crease points. Since the reconstruction of crease discontinuities is particularly relevant in those computer vision problems which require the reconstruction of visible surfaces from two-dimensional images, we have then introduced in (1.2) the function $\Phi(x,u)$. A suitable choice of this function will allow us to apply this variational method to computer vision problems as, for instance, the computation of depth from pairs of stereo images (see [25]).

If the conditions (see [9])

$$\beta\le\alpha\le 2\beta \tag{1.3}$$

are satisfied, the existence of minimizers for the functional $F(u,K_0,K_1)$ has been proved, in the case $n=2$ and $\Phi(x,u)=\mu|u-g|^2$, by Carriero, Leaci, and Tomarelli [13] (notice that (1.3) are necessary and sufficient for the lower semicontinuity of $\overline{F}$ with respect to the L^1 convergence). The proof is based on a weak formulation of the problem by setting

$$\overline{F}(u)=\int_\Omega\left(|\nabla^2 u|^2+\Phi(x,u)\right)\,dx+\alpha\mathcal{H}^{n-1}(S_u)+\beta\mathcal{H}^{n-1}(S_{\nabla u}\setminus S_u)\ , \tag{1.4}$$

where ∇u denotes an approximate differential, S_u is the discontinuity set of u in an approximate sense, and $S_{\nabla u}$ is the discontinuity set of ∇u. In [12] the existence of minimizers for the functional $\overline{F}$ over the space

$$\left\{u:\Omega\to\mathbf{R}:\ u\in L^2(\Omega),\ u\in GSBV(\Omega),\ \nabla u\in[GSBV(\Omega)]^n\right\}\ , \tag{1.5}$$

has been proved in any space dimension n, $GSBV(\Omega)$ being the space of generalized special functions of bounded variation introduced in [17]. A regularity theorem in [13] then shows that, for $n=2$, any weak minimizer actually provides a minimizing triplet (u,K_0,K_1) of F by taking a suitable representative of the function and the closure of S_u and $S_{\nabla u}$.

Ambrosio and Tortorelli [5, 6] approximated the Mumford and Shah functional (1.1) by a family of elliptic functionals defined on Sobolev spaces. The approximation takes place in a variational sense, the De Giorgi Γ-convergence. The approximating elliptic functionals proposed in [6] are defined by

$$E_\epsilon(u,s)=\int_\Omega(s^2+\lambda_\epsilon)|\nabla u|^2\,dx+\mu\int_\Omega|u-g|^2\,dx+\alpha\mathcal{G}_\epsilon(s)\ , \tag{1.6}$$

where the approximation takes place as $\epsilon \to 0^+$, $\lambda_\epsilon \to 0^+$, and

$$\mathcal{G}_\epsilon(s) = \int_\Omega \left[\epsilon |\nabla s|^2 + \frac{(s-1)^2}{4\epsilon} \right] dx \ . \tag{1.7}$$

The variable $s \in [0,1]$ is related to the set of jumps K. The minimizing s_ϵ are near to 0 in a neighborhood of the set K, and far from the neighborhood they are close to 1. The neighborhood shrinks as $\epsilon \to 0$. The Ambrosio and Tortorelli approximation can be used to find an effective algorithm for computing the minimizers of E [25, 29, 30]. The approximation has been applied to several computer vision problems in [28, 32, 33], and further improvements have been proposed and experimented in [34].

In the present paper we consider the following family of functionals:

$$\begin{aligned} F_\epsilon(u,s,\sigma) = &\int_\Omega (\sigma^2 + \kappa_\epsilon) |\nabla^2 u|^2 \, dx + \int_\Omega \Phi(x,u) \, dx + (\alpha - \beta) \mathcal{G}_\epsilon(s) \\ &+ \beta \mathcal{G}_\epsilon(\sigma) + \xi_\epsilon \int_\Omega (s^2 + \zeta_\epsilon) |\nabla u|^\gamma \, dx \end{aligned} \tag{1.8}$$

for suitable infinitesimals κ_ϵ, ξ_ϵ, ζ_ϵ, and $\gamma \geq 2$. A slight variant of these functionals has been proposed by Bellettini and Coscia [7] in the case $n = 1$ and in that case the Γ-convergence of F_ϵ to $\overline{F}$ has been proved (see also the discussion in the beginning of section 6). We extend their Γ-convergence result in the following way: we prove the lower inequality of Γ-convergence in any space dimension n, and we prove the upper inequality when u is bounded and $|\nabla u| \in L^\gamma(\Omega)$, under a very mild regularity assumption on the sets S_u and $S_{\nabla u}$, which is fulfilled in computer vision applications. In the particular case when $\alpha = \beta$ and $n = 2$, we obtain a full Γ-convergence theorem.

The extension of the Ambrosio and Tortorelli approximation to the second order problem presents several difficulties. The lower inequality cannot be obtained by means of the slicing technique and consequent reduction to a one-dimensional problem used in [5, 6]. Such a reduction yields the operator norm of the Hessian matrix in the Γ-limit instead of the euclidean norm. The second derivatives are then estimated by adapting a global technique proposed by Ambrosio in [3] and relying on a compactness theorem in the space (1.5) due to Carriero, Leaci, and Tomarelli [12]. Conversely, the jump part of the functional is estimated by using a slicing argument, taking into account that the space $GSBV$ is a vector space under a suitable energy condition (Proposition 4.3).

The major difficulty in the proof of the upper inequality consists in obtaining a suitable estimate on $\int |\nabla u|^\gamma dx$ from the finiteness of (1.4). Such an estimate would permit us to adapt the constructive part of Ambrosio and Tortorelli's proof [6] to the second order problem. In the case $\alpha = \beta$, $n = \gamma = 2$, an estimate which yields a full Γ-convergence result is obtained by means of a suitable interpolation inequality in $W^{2,2}$ (Proposition 4.6). If $\alpha \neq \beta$, we obtain only a partial result, proving the upper inequality under some mild regularity assumptions on u.

The discretization of the functional (1.2) is not straightforward and it is difficult to apply gradient descent with respect to the unknown sets K_0 and K_1. Conversely, the Γ-convergent approximation yields a sequence of functionals (1.8) which are numerically much more tractable, so that discretization and gradient descent may be applied in a straightforward way. In particular, a simple discretization method, commonly used for computer vision problems [35], may be applied to the functionals (1.8). We then apply the Γ-convergence result to the problem of computation of depth from stereo images, and we present some computer experiments on synthetic images to illustrate the feasibility of the approximation.

2. Notations and preliminary results. Let $\Omega \subset \mathbf{R}^n$ be a bounded open set. We denote by $\mathcal{B}(\Omega)$ the σ-algebra of all the Borel subsets of Ω; for any $C \in \mathcal{B}(\Omega)$ we denote by $\mathrm{meas}(C)$ the Lebesgue n-dimensional measure of C and by $\mathcal{H}^{n-1}(C)$ the Hausdorff $(n-1)$-dimensional measure of C. We denote by $B_\rho(x)$ the open ball $\{y \in \mathbf{R}^n : |y-x| < \rho\}$. We denote by $\mathbf{M}^{n\times n}$ the space of $n \times n$ matrices endowed with the euclidean norm. We introduce the following notations: $s \wedge t = \min\{s,t\}$, $s \vee t = \max\{s,t\}$ for every $s,t \in \mathbf{R}$; given two vectors a, b, we set $\langle a,b\rangle = a \cdot b = \sum_i a_i b_i$ and $(a \otimes b)_{ij} = a_i b_j$.

For any Borel function $u : \Omega \to \mathbf{R}$ we define the approximate upper and lower limits $u^+(x)$, $u^-(x)$ by

$$u^+(x) = \inf\left\{ t \in [-\infty,+\infty] : \lim_{\rho\to 0^+} \frac{\mathrm{meas}(\{y \in B_\rho(x) : u(y) > t\})}{\rho^n} = 0 \right\},$$

$$u^-(x) = \sup\left\{ t \in [-\infty,+\infty] : \lim_{\rho\to 0^+} \frac{\mathrm{meas}(\{y \in B_\rho(x) : u(y) < t\})}{\rho^n} = 0 \right\}.$$

The set

$$S_u = \{x \in \Omega :\ u^-(x) < u^+(x)\}$$

is the discontinuity set of u in an approximate sense and it is negligible with respect to Lebesgue measure (see [20, section 2.9.13]). Suppose $z = u^+(x) = u^-(x) \in \mathbf{R}$; we say that $\nabla u(x) \in \mathbf{R}^n$ is the approximate differential of u at x if $v^+(x) = 0$, where

$$v(y) = \frac{|u(y) - z - \langle \nabla u(x), y-x\rangle|}{|y-x|} \qquad \forall y \in \Omega \setminus \{x\}\ .$$

If u is differentiable at x, then $\nabla u(x)$ is the classical gradient. In the one-dimensional case we shall use the notation u' in place of ∇u. An important property of the approximate differential is the fact that

$$\nabla u(x) = 0 \qquad \text{almost everywhere (a.e.) on} \quad \{y \in \Omega : u(y) = c\} \qquad \forall c \in \mathbf{R}\ . \tag{2.1}$$

We denote by $BV(\Omega)$ the space of functions of bounded variation in Ω, i.e., the functions $u \in L^1(\Omega)$ such that the distributional derivative of u is representable by means of a vector measure $Du = (D_1u, \dots, D_nu)$ with finite total variation. We denote by $|Du|$ the measure total variation of Du. If $u \in BV(\Omega)$, then ∇u exists a.e. in Ω and coincides a.e. with the Radon–Nikodym derivative of Du with respect to the Lebesgue measure [11]. Moreover, the set S_u is countably $(n-1)$-rectifiable, i.e., representable as a disjoint union $\cup_{i=1}^\infty K_i \cup N$, where $\mathcal{H}^{n-1}(N) = 0$ and K_i are compact sets, each contained in a C^1 hypersurface $\Gamma_i \subset \mathbf{R}^n$ [16].

Let $E \subset \mathcal{B}(\Omega)$; we define

$$P(E,\Omega) = \sup\left\{ \int_E \mathrm{div}\ \phi\, dx : \phi \in C_0^1(\Omega;\mathbf{R}^n), |\phi| \le 1 \right\}.$$

We say that E is a set of finite perimeter in Ω if $P(E,\Omega) < +\infty$. By Riesz's theorem (see [21]), E is a set of finite perimeter if and only if $1_E \in BV(\Omega)$, and $P(E,\Omega) = |D1_E|(\Omega)$.

The following Fleming–Rishel coarea formula (see [21]) establishes an important connection between BV functions and sets of finite perimeter:

$$|Du|(\Omega) = \int_{-\infty}^{+\infty} P(\{x \in \Omega : u(x) > t\}, \Omega)\, dt. \tag{2.2}$$

We say that $u \in BV(\Omega)$ belongs to the space of special functions of bounded variation $SBV(\Omega)$ if

$$|Du|(\Omega) = \int_\Omega |\nabla u|\, dx + \int_{S_u} |u^+ - u^-|\, d\mathcal{H}^{n-1} .$$

Functions like the Cantor–Vitali function, whose derivative is concentrated on Cantor's middle third set, are then excluded by $SBV(\Omega)$ (see [1, 17]).

Given a Borel function $u : \Omega \to \mathbf{R}$ we say that $u \in GSBV(\Omega)$ if (see [2, 17])

$$-N \vee u \wedge N \in SBV_{\rm loc}(\Omega) \qquad \forall N \in \mathbf{N}. \tag{2.3}$$

The jump set of u is given by

$$S_u = \bigcup_{N=1}^{\infty} S_{-N\vee u\wedge N}.$$

Furthermore, if $u \in GSBV(\Omega)$, then S_u is countably $(n-1)$-rectifiable, ∇u exists a.e. in Ω and is given by (see [2])

$$\nabla u = \nabla(-N \vee u \wedge N) \quad \text{a.e. on} \quad \{x \in \Omega : |u| \le N\} \qquad \forall N \in \mathbf{N}.$$

We also set

$$GSBV^2(\Omega) = \{u \in GSBV(\Omega) :\ \nabla u \in [GSBV(\Omega)]^n\} .$$

Given $u \in GSBV^2(\Omega)$, we use the notation $\nabla^2_{i,j} u = \nabla_j(\nabla_i u)$ and, in the one-dimensional case, $u'' = (u')'$. Moreover we set

$$S_{\nabla u} = \bigcup_{i=1}^{n} S_{\nabla_i u}.$$

The following compactness result has been proved by Carriero, Leaci, and Tomarelli in [12].

THEOREM 2.1. *Let $(u_h) \subset GSBV^2(\Omega)$ be a sequence such that*

$$\|u_h\|_{L^2} , \qquad \mathcal{H}^{n-1}(S_{u_h} \cup S_{\nabla u_h}) , \qquad \int_\Omega |\nabla^2 u_h|^2\, dx$$

are uniformly bounded in h. Then there exist a subsequence (u_{h_k}) and $u \in GSBV^2(\Omega) \cap L^2(\Omega)$ such that, as $k \to +\infty$,

$$\begin{aligned} &u_{h_k} \to u \qquad \textit{strongly in } L^1(\Omega),\\ &\nabla u_{h_k} \to \nabla u \qquad \textit{a.e. in } \Omega,\\ &\nabla^2 u_{h_k} \rightharpoonup \nabla^2 u \qquad \textit{weakly in } L^2(\Omega; \mathbf{M}^{n\times n}) . \end{aligned}$$

Finally, we recall the following lemma (see [10]).

LEMMA 2.2. *Let $\mu : \mathcal{B}(\Omega) \to [0, +\infty]$ be a σ-finite measure, and let $(f_i) \subset L^1(\Omega)$ be a sequence of nonnegative functions. Then,*

$$\begin{aligned} &\int_\Omega \sup_{i\in\mathbf{N}} f_i(x)\, d\mu(x)\\ &\quad = \sup\left\{ \sum_{i=1}^{k} \int_{A_i} f_i(x)\, d\mu(x) : A_i \subset \Omega \textit{ open and mutually disjoint, } k \in \mathbf{N} \right\}. \end{aligned}$$

We now recall the definition and some properties of Γ-convergence (see [15]). Let X be a metric space and let $f_\epsilon : X \to [0, +\infty]$ be a family of functions indexed by $\epsilon > 0$. We say that f_ϵ Γ-converge as $\epsilon \to 0^+$ to $f : X \to [0, +\infty]$ if the following two conditions

$$\forall x_\epsilon \to x \quad \liminf_{\epsilon \to 0^+} f_\epsilon(x_\epsilon) \geq f(x) \tag{2.4}$$

and

$$\exists x_\epsilon \to x \quad \limsup_{\epsilon \to 0^+} f_\epsilon(x_\epsilon) \leq f(x) \tag{2.5}$$

are fulfilled for every $x \in X$. The Γ-limit, if it exists, is unique and lower semicontinuous. The Γ-convergence is stable under continuous perturbations, that is, $f_\epsilon + g$ Γ-converge to $f + g$ if f_ϵ Γ-converge to f and g is continuous. The most important property of Γ-convergence is the following: if (x_ϵ) is asymptotically minimizing, i.e.,

$$\lim_{\epsilon \to 0^+} \left(f_\epsilon(x_\epsilon) - \inf_X f_\epsilon \right) = 0, \tag{2.6}$$

and if x_{ϵ_h} converge to x for some sequence $\epsilon_h \to 0$, then x minimizes f.

3. Statement of main results. Let $\Phi(x, u) = \mu|u - g|^2$, $\mu > 0$, and $0 < \beta \leq \alpha \leq 2\beta$. For every $u \in GSBV^2(\Omega) \cap L^2(\Omega)$ and every $g \in L^\infty(\Omega)$, we write (1.4) as

$$\overline{F}(u) = \int_\Omega \left(|\nabla^2 u|^2 + \mu|u - g|^2 \right) dx + (\alpha - \beta)\mathcal{H}^{n-1}(S_u) + \beta\mathcal{H}^{n-1}(S_u \cup S_{\nabla u}).$$

In [12], using Theorem 2.1 and a suitable lower semicontinuity theorem in $GSBV^2(\Omega)$, Carriero, Leaci, and Tomarelli proved that the problem

$$(\mathcal{P}) \qquad \min \left\{ \overline{F}(u) : \ u \in GSBV^2(\Omega) \cap L^2(\Omega) \right\}$$

has at least one solution.

For every $\epsilon > 0$ and any function $v \in W^{1,2}(\Omega; [0, 1])$, let us define

$$\mathcal{G}_\epsilon(v) = \int_\Omega \left(\epsilon|\nabla v|^2 + \frac{(v-1)^2}{4\epsilon} \right) dx.$$

Our aim is to approximate $\overline{F}$, in the sense of Γ-convergence, by a family of elliptic functionals F_ϵ which are formally defined by

$$\begin{aligned} F_\epsilon(u, s, \sigma) = &\int_\Omega (\sigma^2 + \kappa_\epsilon)|\nabla^2 u|^2 \, dx + \mu \int_\Omega |u - g|^2 \, dx + (\alpha - \beta)\mathcal{G}_\epsilon(s) \\ &+ \beta\mathcal{G}_\epsilon(\sigma) + \xi_\epsilon \int_\Omega (s^2 + \zeta_\epsilon)|\nabla u|^\gamma \, dx \end{aligned} \tag{3.1}$$

for suitable nonnegative infinitesimals κ_ϵ, ξ_ϵ, ζ_ϵ (in some cases they are allowed to vanish; see the statements below). This formula makes sense if $u \in W^{2,2}(\Omega)$ and s, $\sigma \in W^{1,2}(\Omega)$; however, in the case $\kappa_\epsilon = 0$, because of the coefficient σ^2 multiplying the second derivatives, the functionals F_ϵ are not coercive in these spaces. In section 5 we identify a domain $\mathcal{D}(\Omega)$ of the functionals F_ϵ such that the problem

$$(\mathcal{P}_\epsilon) \qquad \min\{F_\epsilon(u,s,\sigma) : \ (u,s,\sigma) \in \mathcal{D}(\Omega)\}$$

has at least one solution, provided $\gamma > 2$ and $\kappa_\epsilon + \zeta_\epsilon > 0$.

We define

$$X(\Omega) = L^2(\Omega) \times L^\infty(\Omega;[0,1]) \times L^\infty(\Omega;[0,1]) \supset \mathcal{D}(\Omega)$$

and we denote by $\mathcal{F} : X(\Omega) \to [0,+\infty]$ the functional defined by

$$\mathcal{F}(u,s,\sigma) = \begin{cases} \overline{F}(u) & \text{if } u \in GSBV^2(\Omega),\ s \equiv 1,\ \sigma \equiv 1, \\ +\infty & \text{otherwise.} \end{cases}$$

Analogously, we denote by $\mathcal{F}_\epsilon : X(\Omega) \to [0,+\infty]$ the functional defined by

$$\mathcal{F}_\epsilon(u,s,\sigma) = \begin{cases} F_\epsilon(u,s,\sigma) & \text{if } (u,s,\sigma) \in \mathcal{D}(\Omega), \\ +\infty & \text{otherwise.} \end{cases}$$

We first prove the lower inequality of Γ-convergence.

THEOREM 3.1. *Assume that $\gamma \geq 2$, that*

$$\lim_{\epsilon \to 0^+} \frac{\xi_\epsilon}{\epsilon^{\gamma-1}} = +\infty, \tag{3.2}$$

and that either $\kappa_\epsilon > 0$ for ϵ small enough or $\zeta_\epsilon > 0$ for ϵ small enough. Then, for every triple $(u,s,\sigma) \in X(\Omega)$ and for every family $(u_\epsilon, s_\epsilon, \sigma_\epsilon) \in \mathcal{D}(\Omega)$ converging to (u,s,σ) in $[L^1(\Omega)]^3$ as $\epsilon \to 0^+$, we have

$$\liminf_{\epsilon \to 0^+} \mathcal{F}_\epsilon(u_\epsilon, s_\epsilon, \sigma_\epsilon) \geq \mathcal{F}(u,s,\sigma) \ .$$

Moreover, (3.2) can be replaced by the condition $\xi_\epsilon \geq 0$ in the case $\alpha = \beta$.

Then we prove the equicoercivity of the family $(\mathcal{F}_\epsilon)$ under the same assumptions on γ and on the infinitesimals κ_ϵ, ξ_ϵ, ζ_ϵ made in Theorem 3.1.

THEOREM 3.2. *Let $(u_\epsilon, s_\epsilon, \sigma_\epsilon) \in \mathcal{D}(\Omega)$ be such that*

$$\sup_{\epsilon > 0} \mathcal{F}_\epsilon(u_\epsilon, s_\epsilon, \sigma_\epsilon) < +\infty.$$

Then the family $(u_\epsilon, s_\epsilon, \sigma_\epsilon)$ is relatively compact in the $[L^1(\Omega)]^3$ topology as $\epsilon \to 0^+$ and any limit point is of the form $(u,1,1)$ with $u \in GSBV^2(\Omega) \cap L^2(\Omega)$.

We now consider the upper inequality of Γ-convergence. We first state our full Γ-convergence result in the special case when $n = 2$, $\gamma = 2$ and $\alpha = \beta$. We recall that a domain Ω is strictly star-shaped if there exists $x_0 \in \Omega$ such that $t(\Omega - x_0) + x_0 \subset\subset \Omega$ for any $t \in [0,1)$.

THEOREM 3.3. *Assume that $n = \gamma = 2$, $\alpha = \beta$, and Ω is strictly star-shaped. Assume that $\kappa_\epsilon > 0$ and $\kappa_\epsilon = o(\epsilon^4)$, while $\xi_\epsilon = \zeta_\epsilon = 0$. Then the family $(\mathcal{F}_\epsilon)$ Γ-converges to $\mathcal{F}$ in the $[L^1(\Omega)]^3$ topology as $\epsilon \to 0^+$.*

Then from the properties of Γ-convergence and Theorem 3.2, if $(\bar{u}_\epsilon, \bar{s}_\epsilon, \bar{\sigma}_\epsilon)$ minimizes $\mathcal{F}_\epsilon$, then the family $(\bar{u}_\epsilon, \bar{s}_\epsilon, \bar{\sigma}_\epsilon)$ is relatively compact in $[L^1(\Omega)]^3$ as $\epsilon \to 0^+$ and any limit point corresponds to a triple $(u,1,1)$ with u minimizer of $\overline{F}$.

Notice that in the case $\alpha = \beta$, $\xi_\epsilon = 0$, the functionals $\mathcal{F}_\epsilon$ do not depend on s; hence we can write them in the much simpler form

$$F_\epsilon(u,\sigma) = \int_\Omega (\sigma^2 + \kappa_\epsilon)|\nabla^2 u|^2 \, dx + \mu \int_\Omega |u - g|^2 \, dx + \beta \mathcal{G}_\epsilon(\sigma) \ .$$

Now we consider a more general situation. For every set $A \subset \mathbf{R}^n$ and every positive real number ρ, we denote by $(A)_\rho$ the open tubular neighborhood of A with radius ρ, that is, $(A)_\rho = \{x \in \mathbf{R}^n : \ \mathrm{dist}(x, A) < \rho\}$. We define the Minkowski $(n-1)$-dimensional upper and lower content of the set A, respectively, by

$$\mathcal{M}^*(A) = \limsup_{\rho \to 0^+} \frac{\mathrm{meas}((A)_\rho)}{2\rho} \ , \qquad \mathcal{M}_*(A) = \liminf_{\rho \to 0^+} \frac{\mathrm{meas}((A)_\rho)}{2\rho} \ .$$

It can be shown (see [20, section 3.2.39]) that $\mathrm{meas}((A)_\rho)/\rho$ converges to $2\mathcal{H}^{n-1}(A)$ as $\rho \to 0^+$ for any compact subset A of a C^1 hypersurface. In particular, by inner approximation this implies

$$\mathcal{M}_*(A) \geq \mathcal{H}^{n-1}(A)$$

for any $u \in BV(\Omega)$ and any Borel set $A \subset S_u$, because $\mathcal{H}^{n-1}$-almost all of S_u can be covered by C^1 hypersurfaces. The inequality $\mathcal{M}^*(A) \leq \mathcal{H}^{n-1}(A)$, which implies

$$\lim_{\rho \to 0^+} \frac{\mathrm{meas}((A)_\rho)}{2\rho} = \mathcal{H}^{n-1}(A) \ ,$$

holds under very mild regularity assumptions on the set A [5].

We are able to prove the upper inequality of Γ-convergence under the assumption that $u \in L^\infty(\Omega)$, $|\nabla u| \in L^\gamma(\Omega)$ and that, for the sets S_u and $S_u \cup S_{\nabla u}$, Hausdorff measure and Minkowski content coincide.

THEOREM 3.4. *Assume that $\gamma \geq 2$, $\kappa_\epsilon = 0$, $\zeta_\epsilon > 0$, ξ_ϵ satisfies* (3.2) *and $\xi_\epsilon \zeta_\epsilon = o(\epsilon^{\gamma-1})$. Then, for every triple $(u, s, \sigma) \in X(\Omega)$ such that $u \in L^\infty(\Omega)$, $|\nabla u| \in L^\gamma(\Omega)$, and*

$$\mathcal{M}^*(S_u) \leq \mathcal{H}^{n-1}(S_u) \ , \qquad \mathcal{M}^*(S_u \cup S_{\nabla u}) \leq \mathcal{H}^{n-1}(S_u \cup S_{\nabla u}) \ ,$$

there exist $(u_\epsilon, s_\epsilon, \sigma_\epsilon) \in \mathcal{D}(\Omega)$ converging to (u, s, σ) in $[L^1(\Omega)]^3$ as $\epsilon \to 0^+$ such that

$$\limsup_{\epsilon \to 0^+} \mathcal{F}_\epsilon(u_\epsilon, s_\epsilon, \sigma_\epsilon) \leq \mathcal{F}(u, s, \sigma). \tag{3.3}$$

Remark 3.5. The Γ-convergence result still holds if the term $\mu|u - g|^2$ in the functional $\mathcal{F}$ is replaced by $\Phi(x, u)$ in such a way that the functional $u \to \int_\Omega \Phi(x, u)\, dx$ is lower semicontinuous with respect to the strong $L^1(\Omega)$ topology and continuous with respect to the strong $L^2(\Omega)$ topology (see section 7). Let Φ be a Carathéodory function on $\Omega \times \mathbf{R}$, i.e., $\Phi(\cdot, p)$ is measurable for any $p \in \mathbf{R}$ and $\Phi(x, \cdot)$ is continuous for almost every $x \in \Omega$. Then a sufficient condition for Γ-convergence is the following [19]:

$$\begin{cases} \Phi : \Omega \times \mathbf{R} \to \mathbf{R} \quad \textit{is Carathéodory} \ , \\ 0 \leq \Phi(x, u) \leq a(x) + b|u|^2 \ , \end{cases}$$

with $a \in L^1(\Omega)$ and $b \geq 0$.

4. Basic properties of $GSBV^2$ functions. In this section we give some technical results concerning the one-dimensional sections of functions $u \in GSBV(\Omega)$. Let $\nu \in \mathbf{S}^{n-1} = \{x \in \mathbf{R}^n : |x| = 1\}$ be a fixed direction. We set

$$\begin{aligned} &\Pi_\nu = \{x \in \mathbf{R}^n : \langle x, \nu \rangle = 0\}, \\ &\Omega_x = \{t \in \mathbf{R} : x + t\nu \in \Omega\} \quad (x \in \Pi_\nu), \\ &\Omega_\nu = \{x \in \Pi_\nu : \Omega_x \neq \emptyset\}. \end{aligned}$$

The sets Ω_x are the 1-dimensional slices of Ω indexed by $x \in \Pi_\nu$, and Ω_ν is the projection of Ω on Π_ν. Given $u \in GSBV(\Omega)$, we define for $\mathcal{H}^{n-1}$-a.e. $x \in \Omega_\nu$ the restriction

$$u_x(t) = u(x + t\nu) \quad \text{for a.e. } t \in \Omega_x.$$

The following slicing result can be obtained from [1, Theorem 3.3] and [2, section 1].

LEMMA 4.1. *Let $u : \Omega \to \mathbf{R}$ be a measurable function. Then $u \in GSBV(\Omega)$ if and only if, for any $\nu \in \mathbf{S}^{n-1}$, $u_x \in GSBV(\Omega_x)$ for $\mathcal{H}^{n-1}$-a.e. $x \in \Omega_\nu$ and*

$$\int_{A_\nu} |D(-N \vee u_x \wedge N)|(A_x)\, d\mathcal{H}^{n-1} < +\infty \tag{4.1}$$

for any open set $A \subset\subset \Omega$ and any $N \in \mathbf{N}$.

Moreover, if $u \in GSBV(\Omega)$ and $\nu \in \mathbf{S}^{n-1}$, then for $\mathcal{H}^{n-1}$-a.e. $x \in \Omega_\nu$ we have

(a) *$u_x'(t) = \langle \nabla u(x + t\nu), \nu \rangle$ for a.e. $t \in \Omega_x$;*

(b) *$S_{u_x} = (S_u)_x$.*

The proof of the following lemma can be found in Federer [20, section 3.2.22].

LEMMA 4.2. *For every countably $\mathcal{H}^{n-1}$-rectifiable set $E \subset \mathbf{R}^n$ there exists a Borel function $\nu_E : E \to \mathbf{S}^{n-1}$ such that*

$$\int_E |\langle \nu, \nu_E(x) \rangle|\, d\mathcal{H}^{n-1}(x) = \int_{E_\nu} \mathcal{H}^0(E_x)\, d\mathcal{H}^{n-1}(x) \qquad \forall \nu \in \mathbf{S}^{n-1}.$$

The function $\nu_E(x)$ is a normal unit vector to E at x in an approximate sense (see [20, section 3.2.16]).

Although $GSBV(\Omega)$ is not a vector space, we can prove that the natural energy spaces for our problems do have a vector structure.

PROPOSITION 4.3. *The set*

$$Y = \left\{ u \in GSBV(\Omega) : \int_\Omega |\nabla u|\, dx + \mathcal{H}^{n-1}(S_u) < +\infty \right\}$$

is a vector space.

Proof. Let u_1, $u_2 \in Y$, and $\nu \in \mathbf{S}^{n-1}$ be fixed. By Lemma 4.1(a), (b) and Lemma 4.2 we have

$$\int_{\Omega_x} |u_{ix}'|\, dt + \mathcal{H}^0(S_{u_{ix}}) < +\infty \qquad \text{for } i = 1, 2$$

for $\mathcal{H}^{n-1}$-a.e. $x \in \Omega_\nu$, because

$$\int_{\Omega_\nu} \left[\int_{\Omega_x} |u_{ix}'|\, dt + \mathcal{H}^0(S_{u_{ix}}) \right] d\mathcal{H}^{n-1} \le \int_\Omega |\nabla u_i|\, dx + \mathcal{H}^{n-1}(S_{u_i}) < +\infty.$$

In particular, $u_{ix} \in L^\infty_{\mathrm{loc}}(\Omega_x)$, and since $SBV_{\mathrm{loc}}(\Omega_x)$ is a vector space $u_{1x} + u_{2x}$ belongs to $SBV_{\mathrm{loc}}(\Omega_x)$ for $\mathcal{H}^{n-1}$-a.e. $x \in \Omega_\nu$. Since the condition (4.1) is easily verified the conclusion follows by using Lemma 4.1. □

Finally, we show how in $GSBV^2(\Omega)$ second order derivatives and jump set of the derivative can be recovered as well by a slicing method.

LEMMA 4.4. *Let $u \in GSBV^2(\Omega)$ be such that*

$$\int_\Omega |\nabla^2 u|\, dx + \mathcal{H}^{n-1}(S_{\nabla u}) < +\infty.$$

Then, for any $\nu \in \mathbf{S}^{n-1}$ the function $\langle \nabla u, \nu \rangle$ belongs to $GSBV(\Omega)$ and for $\mathcal{H}^{n-1}$-a.e. $x \in \Omega_\nu$ we have

(a) $u'_x \in GSBV(\Omega_x)$;

(b) $u''_x(t) = \langle \nabla \langle \nabla u, \nu \rangle (x + t\nu), \nu \rangle$ *for a.e.* $t \in \Omega_x$;

(c) $S_{u'_x} = (S_{\nabla u \cdot \nu})_x$.

Proof. By Proposition 4.3 it follows that $\langle \nabla u, \nu \rangle \in GSBV(\Omega)$ whenever $\nabla_i u \in GSBV(\Omega)$ for $i = 1, \dots, n$. By Lemma 4.1(a) it follows that $u'_x = \langle \nabla u, \nu \rangle_x$ a.e. in Ω_x for $\mathcal{H}^{n-1}$-a.e. $x \in \Omega_\nu$; in particular, $u'_x \in GSBV(\Omega_x)$ for $\mathcal{H}^{n-1}$-a.e. $x \in \Omega_\nu$. Then, statements (b), (c) follow by applying Lemma 4.1(a,b) to $\langle \nabla u, \nu \rangle$. ∎

COROLLARY 4.5. *The set*

$$\left\{ u \in GSBV^2(\Omega) : \int_\Omega |\nabla^2 u| \, dx + \mathcal{H}^{n-1}(S_u \cup S_{\nabla u}) < +\infty \right\}$$

is a vector space.

Proof. The proof is the same as for Proposition 4.3 using Lemma 4.4 instead of Lemma 4.1. ∎

We conclude this section with an interpolation inequality in $W^{2,2}$ which provides a mild estimate of $\int |\nabla u|^2 \, dx$ with the Blake–Zisserman energy (see also [12]).

PROPOSITION 4.6. *Let A, $B \subset \mathbf{R}^n$ be open sets with $(A)_{2r} \subset\subset B$. Then*

$$\int_A |\nabla u|^2 \, dx \le 16n \left[r^{-2} \int_B u^2 \, dx + 2r^2 \int_B |\nabla^2 u|^2 \, dx \right] \qquad \forall u \in W^{2,2}_{\mathrm{loc}}(B) \ . \tag{4.2}$$

Proof. We prove the inequality only in the case $n = 1$; the general case can be achieved by a slicing argument, taking into account Lemma 4.4(b).

Let x be such that the interval $[x - 2r, x + 2r] \subset B$ and choose $x_1 \in [x + r, x + 2r]$, $x_2 \in [x - 2r, x - r]$ such that

$$r u(x_1) = \int_{x+r}^{x+2r} u(s) \, ds, \qquad r u(x_2) = \int_{x-2r}^{x-r} u(s) \, ds$$

and $x_3 \in [x_2, x_1]$ such that $u'(x_3) = [u(x_1) - u(x_2)]/(x_1 - x_2)$. Then, for any $y \in [x - 2r, x + 2r]$, using twice Hölder inequality we estimate

$$\begin{aligned} |u'(y)|^2 &\le 2|u'(x_3)|^2 + 2 \left(\int_{x_3}^y u''(s) \, ds \right)^2 \\ &\le \frac{4(u^2(x_1) + u^2(x_2))}{r^2} + 2|x_3 - y| \left| \int_{x_3}^y |u''(s)|^2 \, ds \right| \\ &\le \frac{4}{r^3} \int_{x-2r}^{x+2r} u^2(s) \, ds + 8r \int_{x-2r}^{x+2r} |u''(s)|^2 \, ds. \end{aligned}$$

By integration we obtain

$$\int_{x-2r}^{x+2r} |u'|^2 \, dy \le \frac{16}{r^2} \int_{x-2r}^{x+2r} u^2 \, dy + 32r^2 \int_{x-2r}^{x+2r} |u''|^2 \, dy \ .$$

Covering A by a finite number of intervals of length $4r$ contained in B the conclusion follows. ∎

5. The approximation framework. In this section we find a domain suitable for coercivity and lower semicontinuity of the functionals F_ϵ formally defined by (3.1).

We often set $w = (u, s, \sigma)$ and we always assume that $0 \le s \le 1$, $0 \le \sigma \le 1$ almost everywhere. If $\kappa_\epsilon = 0$, we define $p = 2\gamma/(\gamma+2)$ and

$$\mathcal{D}(\Omega) = \left\{(u,s,\sigma) \in X(\Omega) :\ u,\ s,\ \sigma \in W^{1,2}(\Omega),\ \ \sigma\nabla u \in W^{1,p}(\Omega;\mathbf{R}^n)\right\}\ ,$$

if $\kappa_\epsilon > 0$ we define

$$\mathcal{D}(\Omega) = W^{2,2}_{\rm loc}(\Omega) \times W^{1,2}(\Omega;[0,1]) \times W^{1,2}(\Omega;[0,1])\ .$$

If $u \in \mathcal{D}(\Omega)$ and $\kappa_\epsilon = 0$, the approximate differentiability of u and of $\sigma\nabla u$ imply that $\nabla^2 u$ exists a.e. in $\{\sigma > 0\}$ and is given by

$$\nabla^2 u = \frac{\nabla(\sigma\nabla u) - \nabla\sigma \otimes \nabla u}{\sigma}\ . \tag{5.1}$$

We also set $\nabla^2 u = 0$ in $\{\sigma = 0\}$.

In the following we do not need to consider the function s in the case $\alpha = \beta$. We now prove a compactness theorem for the sublevels of F_ϵ.

THEOREM 5.1. *Assume that $\gamma > 2$, $\kappa_\epsilon + \zeta_\epsilon > 0$ and let $(w_h) = (u_h, s_h, \sigma_h) \subset \mathcal{D}(\Omega)$ be a sequence such that*

$$\sup_h F_\epsilon(w_h) < +\infty.$$

Then there exist a subsequence (w_{h_k}) and $w = (u, s, \sigma) \in \mathcal{D}(\Omega)$ such that (w_{h_k}) converge in $[L^1(\Omega)]^3$ to w and (∇u_{h_k}) converge a.e. to ∇u in $\{\sigma > 0\}$.

Proof. From (4.2), in the case $\kappa_\epsilon > 0$, we have that (u_h) is bounded in $W^{2,2}(A)$ for any open set $A \subset\subset \Omega$. The statement then follows from Rellich theorem. Hence, in the following we consider the more delicate case when $\kappa_\epsilon = 0$ and $\zeta_\epsilon > 0$.

From the definition of F_ϵ the sequences (s_h) and (σ_h) are bounded in $W^{1,2}(\Omega)$. Moreover, since $(|\nabla u_h|)$ is bounded in $L^\gamma(\Omega)$ and

$$\nabla(\sigma_h \nabla u_h) = \sigma_h \nabla^2 u_h + \nabla\sigma_h \otimes \nabla u_h,$$

$v_h = \sigma_h \nabla u_h$ are also bounded in $W^{1,p}(\Omega;\mathbf{R}^n)$. Hence, possibly extracting a further subsequence we can assume that (v_{h_k}) is converging a.e. in Ω. It easily follows that $\nabla u_{h_k} = v_{h_k}/\sigma_{h_k}$ converge a.e. to ∇u in $\{\sigma > 0\}$.

In order to prove that $\sigma\nabla u \in W^{1,p}(\Omega;\mathbf{R}^n)$ (hence $w \in \mathcal{D}(\Omega)$), we notice that by Hölder inequality, we have

$$\lim_{h\to+\infty} \int_{\{\sigma=0\}} |\sigma_h \nabla u_h|\, dx = 0$$

hence (possibly extracting a subsequence) $\sigma_h\nabla u_h$ converge a.e. to $\sigma\nabla u$ in the whole of Ω. Since $(\sigma_h\nabla u_h)$ is also bounded in $W^{1,p}(\Omega;\mathbf{R}^n)$, it follows that $\sigma\nabla u \in W^{1,p}(\Omega;\mathbf{R}^n)$ and that $\sigma_{h_k}\nabla u_{h_k}$ weakly converge in $W^{1,p}(\Omega;\mathbf{R}^n)$ to $\sigma\nabla u$. □

Now we prove the lower semicontinuity of F_ϵ.

THEOREM 5.2. *Assume that $\gamma > 2$, $\kappa_\epsilon + \zeta_\epsilon > 0$ and let $(w_h) = (u_h, s_h, \sigma_h) \subset \mathcal{D}(\Omega)$ be converging in $[L^1(\Omega)]^3$ to $w = (u, s, \sigma) \in \mathcal{D}(\Omega)$. Then*

$$\liminf_{h\to+\infty} F_\epsilon(w_h) \ge F_\epsilon(w)\ .$$

Proof. In this case we also consider only the more difficult case when $\kappa_\epsilon = 0$ and $\zeta_\epsilon > 0$. It is not restrictive to assume that $(F_\epsilon(w_h))$ is converging to a finite limit and, by Theorem 5.1 and its proof, we can also assume that ∇u_h converge to ∇u a.e. in $\{\sigma > 0\}$ and $\sigma_h \nabla u_h$ weakly converge in $W^{1,p}(\Omega; \mathbf{R}^n)$ to $\sigma \nabla u$.

Since the sequences (s_h) and (σ_h) are bounded in $W^{1,2}(\Omega)$, they weakly converge, respectively, to s, σ and therefore the terms $\mathcal{G}_\epsilon(s)$ and $\mathcal{G}_\epsilon(\sigma)$ are lower semicontinuous. The lower semicontinuity of $\int_\Omega (s^2 + \zeta_\epsilon)|\nabla u|^\gamma \, dx$ directly follows by Ioffe lower semicontinuity theorem (see [10, Theorem 4.1.1]).

Finally, the identity

$$\nabla(\sigma_h \nabla u_h) = \sigma_h \nabla^2 u_h + \nabla \sigma_h \otimes \nabla u_h$$

and the weak convergence of $\nabla(\sigma_h \nabla u_h)$ to $\nabla(\sigma \nabla u)$ easily imply that $\nabla^2 u_h$ weakly converge to $\nabla^2 u$ in $L^2(K; \mathbf{M}^{n\times n})$ on any compact set $K \subset \Omega$ on which (σ_h) uniformly converges to σ, (∇u_h) uniformly converges to ∇u, and $\inf_K \sigma > 0$. Then, Ioffe lower semicontinuity theorem again gives

$$\int_K \sigma^2 |\nabla^2 u|^2 \, dx \le \liminf_{h\to+\infty} \int_K \sigma_h^2 |\nabla^2 u_h|^2 \, dx.$$

Let $\delta > 0$; by Egorov theorem we can cover almost all of $\{\sigma \ge \delta\}$ by an increasing sequence of compact sets on which (σ_h) and (∇u_h) are uniformly converging. As a consequence, the inequality above holds with $\{\sigma \ge \delta\}$ in place of K, and letting $\delta \downarrow 0$ we obtain the lower semicontinuity of the term $\int_\Omega \sigma^2 |\nabla^2 u|^2 \, dx$. □

From the compactness and the lower semicontinuity properties of the functional F_ϵ it follows that for any $\epsilon > 0$ the problem

$$(\mathcal{P}_\epsilon) \qquad \min \{F_\epsilon(u, s, \sigma) : \ (u, s, \sigma) \in \mathcal{D}(\Omega)\}$$

has at least one solution, provided $\gamma > 2$ and $\kappa_\epsilon + \zeta_\epsilon > 0$. Finally, if $\kappa_\epsilon > 0$, the problem $(\mathcal{P}_\epsilon)$ has a solution also in the case $\gamma = 2$.

6. The lower inequality. In this section we prove the lower inequality of Γ-convergence (2.4) and the equicoercivity of the family $(\mathcal{F}_\epsilon)$. In the following it will be convenient also to consider functionals depending on the domain of integration.

The following lower bound for the jump terms in the one-dimensional case has been shown by Bellettini and Coscia in [7, Theorem 3.1].

LEMMA 6.1. *Assume that κ_ϵ, ξ_ϵ, ζ_ϵ are as in Theorem* 3.1 *and $\gamma \ge 2$. Let $I \subset \mathbf{R}$ be a bounded open set and $\epsilon_h \to 0^+$. Then, for every sequence (w_h) converging to w in $[L^1(I)]^3$ as $h \to +\infty$ such that $\mathcal{F}_{\epsilon_h}(w_h)$ is bounded, we have*

$$\begin{aligned} \liminf_{h\to+\infty} [(\alpha - \beta)\mathcal{G}_{\epsilon_h}(s_h, I) + \beta \mathcal{G}_{\epsilon_h}(\sigma_h, I)] &\ge (\alpha - \beta)\mathcal{H}^0(S_u \cap I) \\ &\quad + \beta \mathcal{H}^0\left((S_u \cup S_{u'}) \cap I\right) \ . \end{aligned} \tag{6.1}$$

The condition (3.2) *on ξ_ϵ can be dropped in the case $\alpha = \beta$.*

Since our functionals are slightly different from those in [7], some remarks are necessary. Indeed, the functionals in [7] are given by

$$\begin{aligned} F_\epsilon(u, s, \sigma) = &\int_\Omega (\sigma^2 + \kappa_\epsilon)|\nabla^2 u|^2 \, dx + \mu \int_\Omega |u - g|^2 \, dx + (\alpha - \beta)\mathcal{G}_\epsilon(s) \\ &+ \beta \mathcal{G}_\epsilon(\sigma) + \xi_\epsilon \int_\Omega s^2 |\nabla u|^2 \, dx \ , \end{aligned}$$

hence the differences with respect to ours are two: first they assume that $\zeta_\epsilon = 0$ and $\gamma = 2$ and then they prove the lower bound only in the case when $\kappa_\epsilon > 0$ (hence $u_h \in W^{2,2}(I)$). The assumption that $\zeta_\epsilon = 0$ is not a problem, since smaller functionals are considered, and also a general exponent γ can be considered, provided (3.2) holds. However, for technical reasons related to the proof of the Γ-limsup inequality, in particular, the difficulty in estimating the second derivatives of $u_\epsilon = \psi_\epsilon u$ in the proof of Theorem 3.4 (this can be avoided in the case $n = 1$ using suitable interpolating cubic polynomials), we have preferred a different formulation of the energy in the larger class $\mathcal{D}(\Omega)$, which still provides compactness of minimizing sequences and lower semicontinuity of the energy. Moreover, the proof of the Γ-liminf inequality of Bellettini and Coscia works, essentially with no modification, also for our more general functionals. Notice also that our full Γ-convergence result Theorem 3.3 fits exactly in the Bellettini and Coscia framework.

The reason why no condition on ξ_ϵ (besides $\xi_\epsilon \geq 0$) is necessary in the case $\alpha = \beta$ is that the term $\xi_\epsilon \int s^2 |\nabla u|^\gamma \, dx$ has been added to the energy to force σ_h to tend to zero at least *twice* (paying asymptotically at least $2\beta \geq \alpha$) close to jumps of u if s_h is far away from 0 (if this does not happen and (3.2) holds, then the additional term diverges; see Lemma 3.2(i) of [7]); in the case when $\alpha = \beta$ it is not necessary to force this behavior of σ_h, since σ_h is already forced by the other terms of F_ϵ to tend to zero at least *once* (paying asymptotically at least β) close to jumps of u or of u', regardless of the values of s_h.

Finally, we notice that we can restate (6.1) as follows:

$$\begin{aligned} &\liminf_{h\to+\infty} \left[t\mathcal{F}_{\epsilon_h}(w_h) + (\alpha-\beta)\mathcal{G}_{\epsilon_h}(s_h, I) + \beta \mathcal{G}_{\epsilon_h}(\sigma_h, I) \right] \\ &\geq (\alpha-\beta)\mathcal{H}^0(S_u \cap I) + \beta \mathcal{H}^0\left((S_u \cup S_{u'}) \cap I \right) \qquad \forall t > 0 \, . \end{aligned} \tag{6.2}$$

The advantage of this new formulation is that the a priori assumption that $\mathcal{F}_{\epsilon_h}(w_h)$ is bounded can be dropped.

6.1. Proof of Theorem 3.1. Let $(w_\epsilon) \in \mathcal{D}(\Omega)$, $w \in X(\Omega)$, be such that $w_\epsilon \to w$ in $[L^1(\Omega)]^3$ as $\epsilon \to 0^+$. We assume that

$$+\infty > L = \liminf_{\epsilon\to 0} \mathcal{F}_\epsilon(w_\epsilon, \Omega) = \lim_{h\to+\infty} \mathcal{F}_{\epsilon_h}(w_{\epsilon_h}, \Omega) \, , \tag{6.3}$$

otherwise the result is trivial. For notational simplicity we set $w_{\epsilon_h} = (u_h, s_h, \sigma_h)$ and we assume that w_{ϵ_h} converge a.e. to (u, s, σ) as $h \to +\infty$.

We also assume that (w_h) converges to w fast enough, i.e., $\sum_h \|w_h - w\|_{L^1} < +\infty$. This assumption and Fubini theorem imply (with the notation of section 4)

$$\lim_{h\to\infty} w_{hx} = w_x \quad \text{a.e. in } \Omega_x \qquad\qquad \text{for } \mathcal{H}^{n-1}\text{-a.e. } x \in \Omega_\nu$$

for any direction $\nu \in \mathbf{S}^{n-1}$, and this will be useful in what follows.

If either s or σ were not identically equal to 1, then by the Fatou's lemma we would get

$$L \geq \liminf_{h\to+\infty} \left[(\alpha-\beta) \int_{\{s\neq 1\}} \frac{(s_h-1)^2}{4\epsilon_h} \, dx + \beta \int_{\{\sigma\neq 1\}} \frac{(\sigma_h-1)^2}{4\epsilon_h} \, dx \right] \geq +\infty,$$

which contradicts the assumption that $L < +\infty$. Therefore, we will assume that $s \equiv 1$ and $\sigma \equiv 1$. As before we do not need to consider the function s in the case $\alpha = \beta$.

The proof now follows by proving separately the following inequalities:

$$\liminf_{h\to+\infty} \int_\Omega {\sigma_h}^2 |\nabla^2 u_h|^2 \, dx \geq \int_\Omega |\nabla^2 u|^2 \, dx \ , \tag{6.4}$$

$$\begin{aligned} \liminf_{h\to+\infty} [(\alpha-\beta)\mathcal{G}_{\epsilon_h}(s_h,\Omega) + \beta \mathcal{G}_{\epsilon_h}(\sigma_h,\Omega)] &\geq (\alpha-\beta)\mathcal{H}^{n-1}(S_u) \\ &\quad + \beta \mathcal{H}^{n-1}(S_u \cup S_{\nabla u}) \ . \end{aligned} \tag{6.5}$$

The lower semicontinuity of the term $\int |u-g|^2 \, dx$ with respect to the strong $L^1(\Omega)$ topology then completes the proof.

Possibly extracting a subsequence (this is allowed, since we are assuming that $\mathcal{F}_{\epsilon_h}(w_{\epsilon_h})$ is converging) we can assume that both lim inf in (6.4) and (6.5) are finite limits, denoted by L_1 and L_2, respectively.

We first prove (6.4). Let $\psi(t) = \int_0^t (1-\tau)\, d\tau$; using (6.3) we have

$$\int_\Omega |\nabla \psi(\sigma_h)| \, dx = \int_\Omega |\nabla \sigma_h| (1-\sigma_h) \, dx \leq \int_\Omega \left[\epsilon_h |\nabla \sigma_h|^2 + \frac{(1-\sigma_h)^2}{4\epsilon_h} \right] dx \leq \frac{L+1}{\beta}$$

for h large enough. Then, by the coarea formula (2.2), we have

$$\int_0^{\psi(1)} P(\{\psi(\sigma_h) > t\}, \Omega) \, dt = \int_\Omega |\nabla \psi(\sigma_h)| \, dx \leq \frac{L+1}{\beta}$$

for h large enough. By the Fatou lemma we then get

$$\int_{\psi(a)}^{\psi(1)} \liminf_{h\to+\infty} P(\{\psi(\sigma_h) > t\}, \Omega) \, dt \leq \liminf_{h\to+\infty} \int_\Omega |\nabla \psi(\sigma_h)| \, dx \leq \frac{L+1}{\beta}$$

for any $a \in (0,1)$. Therefore there exists $t_0 = \psi(\theta) \in (\psi(a), \psi(1))$ for some $\theta \in (a,1)$ such that

$$\liminf_{h\to+\infty} P(\{\psi(\sigma_h) > t_0\}, \Omega) \leq l < +\infty \tag{6.6}$$

with $l = (L+1)/[\beta(\psi(1)-\psi(a))]$.

Then, if we set $E_h = \{\sigma_h > \theta\}$, by (6.6) we get $P(E_h, \Omega) \leq l+1$ for infinitely many h; for notational simplicity we will assume in the following that the inequality is true for any h (in the general case a further subsequence must be extracted). By the L^1 convergence of (σ_h) to 1 we obtain

$$\mathrm{meas}(\Omega \setminus E_h) \leq \frac{1}{1-\theta} \int_\Omega (1-\sigma_h) \, dx \to 0 \ .$$

Then we define

$$v_h = u_h 1_{E_h} \ . \tag{6.7}$$

By the locality property (2.1) we get

$$\nabla v_h = 1_{E_h} \nabla u_h \ , \qquad \nabla^2 v_h = 1_{E_h} \nabla^2 u_h$$

for a.e. $x \in \Omega$. Since

$$-N \vee v_h \wedge N = 1_{E_h} [-N \vee u_h \wedge N] \qquad \forall N \in \mathbf{N} \ ,$$

and taking into account that E_h has finite perimeter, from [36, Chapter 4, section 6.4] it follows that $v_h \in GSBV(\Omega)$. Analogously,

$$-N \vee (\nabla_i v_h) \wedge N = 1_{E_h}[-N \vee (\nabla_i u_h) \wedge N] \in SBV_{\rm loc}(\Omega)$$

for any $N \in \mathbf{N}$ and any $i = 1, \dots, n$. Then $v_h \in GSBV^2(\Omega)$ and we have

$$\mathcal{H}^{n-1}(S_{v_h} \cup S_{\nabla v_h}) \le P(E_h, \Omega) \le l + 1 \quad \text{for every } h \in \mathbf{N} \ .$$

Then, since $\sigma_h \ge \theta$ on E_h, the sequence (v_h) satisfies all the assumptions of the compactness Theorem 2.1, hence we can assume (again, possibly passing to a subsequence) that (v_h) converges in $L^1(\Omega)$ to some function $v \in GSBV^2(\Omega)$ with $\nabla v_h \to \nabla v$ a.e. in Ω and $\nabla^2 v_h$ weakly converging to $\nabla^2 v$ in $L^2(\Omega; \mathbf{M}^{n\times n})$. Since (u_h) converges to u in $L^1(\Omega)$ and $\mathrm{meas}(\Omega \setminus E_h) \to 0$, we obtain that $u = v \in GSBV^2(\Omega)$; moreover, by the lower semicontinuity of quadratic forms with respect to weak convergence in L^2 we get

$$L_1 \ge \liminf_{h\to+\infty} \int_{E_h} \theta^2 |\nabla^2 u_h|^2 \, dx = \liminf_{h\to+\infty} \int_\Omega \theta^2 |\nabla^2 v_h|^2 \, dx \ge \int_\Omega \theta^2 |\nabla^2 u|^2 \, dx \ .$$

By letting $a \uparrow 1$ (hence $\theta \to 1$) we obtain (6.4).

The relation (6.5) will be proved using (6.2) and a slicing argument. Let $A \subset \Omega$ be open and $\nu \in \mathbf{S}^{n-1}$ be fixed. By using the notation of section 4 we have

$$\begin{aligned}
\mathcal{G}_{\epsilon_h}(s_h, A) &\ge \int_A \left(\epsilon_h |\langle \nabla s_h, \nu \rangle|^2 + \frac{(s_h - 1)^2}{4\epsilon_h} \right) dx \\
&= \int_{A_\nu} d\mathcal{H}^{n-1}(x) \int_{A_x} \left(\epsilon_h |s'_{hx}|^2 + \frac{(s_{hx} - 1)^2}{4\epsilon_h} \right) dt \\
&= \int_{A_\nu} \mathcal{G}_{\epsilon_h}(s_{hx}, A_x) \, d\mathcal{H}^{n-1}(x) \ .
\end{aligned}$$

An analogous relation holds for $\mathcal{G}_{\epsilon_h}(\sigma_h, A)$ and, taking into account Lemma 4.4, for $\mathcal{F}_{\epsilon_h}(w_h, A)$.

Since w_{hx} converge to w_x in $[L^1(\Omega_x)]^3$ for $\mathcal{H}^{n-1}$-almost every $x \in \Omega_\nu$, by using Fatou's lemma, (6.2), Lemmas 4.1 and 4.4, and eventually Lemma 4.2, we get

$$\begin{aligned}
&\liminf_{h\to+\infty} [t\mathcal{F}_{\epsilon_h}(w_h) + (\alpha - \beta)\mathcal{G}_{\epsilon_h}(s_h, A) + \beta \mathcal{G}_{\epsilon_h}(\sigma_h, A)] \\
&\ge \int_{A_\nu} \liminf_{h\to+\infty} [t\mathcal{F}_{\epsilon_h}(w_{hx}) + (\alpha - \beta)\mathcal{G}_{\epsilon_h}(s_{hx}, A_x) + \beta \mathcal{G}_{\epsilon_h}(\sigma_{hx}, A_x)] \, d\mathcal{H}^{n-1}(x) \\
&\ge (\alpha - \beta) \int_{A_\nu} \mathcal{H}^0(S_{u_x} \cap A_x) \, d\mathcal{H}^{n-1}(x) + \beta \int_{A_\nu} \mathcal{H}^0((S_{u_x} \cup S_{u'_x}) \cap A_x) \, d\mathcal{H}^{n-1}(x) \\
&= (\alpha - \beta) \int_{A_\nu} \mathcal{H}^0((S_u \cap A)_x) \, d\mathcal{H}^{n-1}(x) + \beta \int_{A_\nu} \mathcal{H}^0(((S_u \cup S_{\nabla u \cdot \nu}) \cap A)_x) \, d\mathcal{H}^{n-1}(x) \\
&= \alpha \int_{S_u \cap A} |\langle \nu, \nu_u(x) \rangle| \, d\mathcal{H}^{n-1}(x) + \beta \int_{(S_{\nabla u \cdot \nu} \setminus S_u) \cap A} |\langle \nu, \nu_{\nabla u}(x) \rangle| \, d\mathcal{H}^{n-1}(x)
\end{aligned}$$

for any $t > 0$, where $\nu_u(x)$ and $\nu_{\nabla u}(x)$ are approximate unit normals to S_u and $S_{\nabla u}$, respectively. Since $\mathcal{F}_{\epsilon_h}(w_{\epsilon_h}, \Omega)$ converges to L, then by letting $t \downarrow 0$ we obtain

$$\begin{aligned}
&\liminf_{h\to+\infty} [(\alpha - \beta)\mathcal{G}_{\epsilon_h}(s_h, A) + \beta \mathcal{G}_{\epsilon_h}(\sigma_h, A)] \\
&\quad \ge \alpha \int_{S_u \cap A} |\langle \nu, \nu_u(x) \rangle| \, d\mathcal{H}^{n-1}(x) + \beta \int_{(S_{\nabla u \cdot \nu} \setminus S_u) \cap A} |\langle \nu, \nu_{\nabla u}(x) \rangle| \, d\mathcal{H}^{n-1}(x) \ .
\end{aligned} \tag{6.8}$$

We now apply Lemma 2.2 in the following framework:

- $f_\nu(x) = \alpha|\langle \nu, \nu_u(x)\rangle| 1_{S_u} + \beta|\langle \nu, \nu_{\nabla u}(x)\rangle| 1_{S_{\nabla u\cdot\nu}\setminus S_u}$,
- $\mu = \mathcal{H}^{n-1}\lfloor(S_u \cup S_{\nabla u})$,

with ν varying in a countable dense subset D of $\mathbf{S}^{n-1}$. Since $\sup_{\nu\in D} f_\nu = \alpha 1_{S_u} + \beta 1_{S_{\nabla u}\setminus S_u}$ (because any $x \in S_{\nabla u}$ belongs to $S_{\nabla u\cdot\nu}$ provided $\langle \nu, \nabla^+ u(x) - \nabla^- u(x)\rangle \neq 0$), by Lemma 2.2 we have that

$$(\alpha-\beta)\mathcal{H}^{n-1}(S_u) + \beta\mathcal{H}^{n-1}(S_u \cup S_{\nabla u})$$

is equal to the supremum of

$$\sum_{i=1}^k \left\{ \alpha \int_{S_u\cap A_i} |\langle \nu_i, \nu_u(x)\rangle|\, d\mathcal{H}^{n-1}(x) + \beta \int_{(S_{\nabla u\cdot\nu_i}\setminus S_u)\cap A_i} |\langle \nu_i, \nu_{\nabla u}(x)\rangle|\, d\mathcal{H}^{n-1}(x) \right\}$$

among all finite families (A_i, ν_i) with $\nu_i \in D$ and $A_i \subset \Omega$ open and pairwise disjoint. By (6.8) and the superadditivity of the lim inf operator, any of these sums is less than L_2, whence the inequality (6.5) follows (see also [5]). □

6.2. Proof of Theorem 3.2. By the equiboundedness of $\mathcal{F}_\epsilon(w_\epsilon)$ it follows as before that $(s_\epsilon, \sigma_\epsilon) \to (1,1)$ in $[L^1(\Omega)]^2$ as $\epsilon \to 0^+$.

Reasoning as in the proof of (6.4) of Theorem 3.1 we can find a sequence $\epsilon_h \to 0^+$ and measurable sets E_h such that $\mathrm{meas}(\Omega \setminus E_h) \to 0$ and $v_{\epsilon_h} = u_{\epsilon_h} 1_{E_h}$ converge in $L^1(\Omega)$ to $u \in GSBV^2(\Omega) \cap L^2(\Omega)$. Since (u_ϵ) is equibounded in $L^2(\Omega)$, by Hölder inequality $\|u_{\epsilon_h} - v_{\epsilon_h}\|_{L^1} \to 0$ as $h \to +\infty$, hence $u_{\epsilon_h} \to u$ in $L^1(\Omega)$. □

7. The upper inequality.

7.1. Proof of Theorem 3.3. We can assume without losing generality that $u \in GSBV^2(\Omega) \cap L^2(\Omega)$, $|\nabla^2 u| \in L^2(\Omega)$, and $\sigma \equiv 1$. Since we are assuming that $\alpha = \beta$ we simply set $s_\epsilon \equiv 1$ for any $\epsilon > 0$. We construct a family u_ϵ converging to u in $L^2(\Omega)$, so that we can neglect the term $\mu \int |u-g|^2\, dx$, which is continuous with respect to the strong $L^2(\Omega)$ topology, and we then assume $\mu = 0$.

Assuming that Ω is star-shaped with respect to the origin, we set $\Omega_t = t\Omega$ with $t \in (0,1)$ and construct a family $w_\epsilon = (u_\epsilon, s_\epsilon, \sigma_\epsilon) \in \mathcal{D}(\Omega_t)$ such that (as in the previous section we emphasize the dependence on the domain of integration)

$$\limsup_{\epsilon\to 0^+} \mathcal{F}_\epsilon(w_\epsilon, \Omega_t) \le \mathcal{F}(w, \Omega)\ . \tag{7.1}$$

Then, the functions $w_{\epsilon,t}(x) = w_\epsilon(tx)$ belong to $\mathcal{D}(\Omega)$ and satisfy

$$\limsup_{\epsilon\to 0^+} \mathcal{F}_\epsilon(w_{\epsilon,t}, \Omega) \le t^{-n}\mathcal{F}(w, \Omega)\ ,$$

hence the desired family of the Γ-limsup inequality can be constructed by a diagonal argument by letting $t \uparrow 1$.

In order to construct the family (w_ϵ) satisfying (7.1) we follow the outline of [6], assuming first that

$$\mathcal{M}^*\left((S_u \cup S_{\nabla u}) \cap K\right) = \mathcal{H}^{n-1}\left((S_u \cup S_{\nabla u}) \cap K\right) \text{ for any } K \subset \Omega \text{ compact}\ . \tag{7.2}$$

We restrict our choice to the functions u_ϵ and σ_ϵ that, outside a tubular neighborhood of $S_u \cup S_{\nabla u}$, with radius depending on ϵ, are, respectively, equal to u and 1.

Setting $\tilde{S}_u = S_u \cup S_{\nabla u}$ and $\tau(x) = \operatorname{dist}(x, \tilde{S}_u)$, by the interpolation inequality (4.2) we obtain a constant C depending only on u such that

$$\int_{\Omega_t \setminus (\tilde{S}_u)_r} |\nabla u|^2 \, dx \le C r^{-2} \tag{7.3}$$

for any r sufficiently small. In view of our assumption on κ_ϵ, we can find an infinitesimal b_ϵ faster than ϵ such that $\kappa_\epsilon = o(b_\epsilon^4)$ (for instance, $b_\epsilon = (\epsilon \kappa_\epsilon^{1/4})^{1/2}$), and an infinitesimal η_ϵ faster than $\sqrt{\epsilon}$ such that $a_\epsilon = -2\epsilon \ln \eta_\epsilon$ is infinitesimal (for instance, $\eta_\epsilon = \epsilon$).

With this choice of infinitesimals, we then define

$$\sigma_\epsilon(x) = \begin{cases} 0 & \text{if } x \in (\tilde{S}_u)_{b_\epsilon}, \\ 1 - \eta_\epsilon & \text{if } x \in \Omega_t \setminus (\tilde{S}_u)_{a_\epsilon + b_\epsilon} \ . \end{cases}$$

Let now y_ϵ be the solution of the Cauchy problem

$$\dot{y}(t) = \frac{1 - y}{2\epsilon}, \qquad y(b_\epsilon) = 0,$$

that is, $y_\epsilon(t) = 1 - \exp\left[(b_\epsilon - t)/(2\epsilon)\right]$. We complete the definition of σ_ϵ by setting

$$\sigma_\epsilon(x) = y_\epsilon \circ \tau(x) \qquad \text{if } x \in (\tilde{S}_u)_{a_\epsilon + b_\epsilon} \setminus (\tilde{S}_u)_{b_\epsilon} \ .$$

Now we turn to the choice of u_ϵ. To this aim, we build a smooth function $\psi_\epsilon : \Omega \to [0,1]$ such that $\psi_\epsilon = 0$ in $\{\tau \le b_\epsilon/2\}$, $\psi_\epsilon = 1$ in $\{\tau \ge b_\epsilon\}$, and $|\nabla \psi_\epsilon| = O(1/b_\epsilon)$, $|\nabla^2 \psi_\epsilon| = O(1/b_\epsilon^2)$. Taking into account that $|\nabla \tau| = 1$ a.e., a function ψ_ϵ with the required properties can be built as $(\chi \circ \tau) * \rho$, where ρ is a convolution kernel with diameter $b_\epsilon/3$ and $\chi(s) = [0 \vee (6s/b_\epsilon - 4) \wedge 1]$. The assumptions on u and the interpolation inequality (4.2) yield

$$u \in W^{2,2}_{\mathrm{loc}}(\Omega \setminus \overline{\tilde{S}_u}) \ ,$$

so that, if we set $u_\epsilon = u\psi_\epsilon$, we have $u_\epsilon \in W^{2,2}(\Omega_t)$.

With these choices, we get

$$\mathcal{F}_\epsilon(w_\epsilon, \Omega_t) = \int_{\Omega_t} (\sigma_\epsilon^2 + \kappa_\epsilon) |\nabla^2 u_\epsilon|^2 \, dx \tag{7.4}$$

$$+\beta \mathcal{G}_\epsilon(\sigma_\epsilon, \Omega_t \cap ((\tilde{S}_u)_{a_\epsilon + b_\epsilon} \setminus (\tilde{S}_u)_{b_\epsilon})) \tag{7.5}$$

$$+\beta \frac{\operatorname{meas}(\Omega_t \cap (\tilde{S}_u)_{b_\epsilon})}{4\epsilon} \tag{7.6}$$

$$+\beta \frac{\eta_\epsilon^2}{4\epsilon} \operatorname{meas}(\Omega_t \setminus (\tilde{S}_u)_{a_\epsilon + b_\epsilon}) \ . \tag{7.7}$$

Since $u_\epsilon \equiv u$ on $\{\sigma_\epsilon > 0\}$ (because $\psi_\epsilon \equiv 1$ on $\{\tau \ge b_\epsilon\} \supset \{\sigma_\epsilon > 0\}$), the upper limit of the term in (7.4) does not exceed

$$\int_\Omega |\nabla^2 u|^2 \, dx + \limsup_{\epsilon \to 0^+} \kappa_\epsilon \int_{\Omega_t \cap \{b_\epsilon/2 \le \tau \le b_\epsilon\}} |\nabla^2 u_\epsilon|^2 \, dx \ .$$

Taking into account (7.3), the identity

$$\nabla^2 u_\epsilon = \psi_\epsilon \nabla^2 u + 2\nabla\psi_\epsilon \otimes \nabla u + u\nabla^2\psi_\epsilon \ ,$$

and our choice of b_ϵ we obtain that the lim sup above is zero.

Concerning the term in (7.5), from the proof of Theorem 3.1 of [6] it follows that its upper limit does not exceed $\beta\mathcal{M}^*(\tilde{S}_u \cap \overline{\Omega}_t)$; by (7.2) we obtain that the upper limit is less than $\beta\mathcal{H}^{n-1}(\tilde{S}_u)$.

The term in (7.6) is infinitesimal because the Minkowski content is finite and $b_\epsilon = o(\epsilon)$, and similarly the term in (7.7) is infinitesimal because $\eta_\epsilon^2 = o(\epsilon)$.

This proves, under the additional assumption (7.2), the existence of a family (w_ϵ) satisfying (7.1). The assumption can be removed as follows. Consider for any $\lambda > 0$ the penalized problem

$$\min\left\{\int_\Omega \left(|\nabla^2 v|^2 + \lambda|v-u|^2\right)\, dx + \beta\mathcal{H}^{n-1}\left(S_v \cup S_{\nabla v}\right) :\ v \in GSBV^2(\Omega) \cap L^2(\Omega)\right\}\ ,$$

and let u_λ be a minimizer (see [12]). Notice that $\overline{F}(u_\lambda) \le \overline{F}(u) < +\infty$, hence $u_\lambda \to u$ as $\lambda \to +\infty$. Then, it has been proved in [14] that any function u_λ fulfills (7.2), and therefore a family (w_ϵ) satisfying (7.1) for $(u,1,1)$ can be obtained from those already constructed for $(u_\lambda,1,1)$ by a diagonal argument. □

7.2. Proof of Theorem 3.4. Since the proof is similar to that of Theorem 3.3 we sketch only the relevant differences. The function σ_ϵ is defined in the same way and s_ϵ is constructed analogously in a tubular neighborhood of S_u. Let $\tau_1(x) = \mathrm{dist}(x, S_u)$. In order to construct u_ϵ we fix some smooth function ψ_ϵ such that $\psi_\epsilon = 0$ in $\{\tau_1 \le b_\epsilon/2\}$, $\psi_\epsilon = 1$ in $\{\tau_1 \ge b_\epsilon\}$, and $|\nabla\psi_\epsilon| = O(1/b_\epsilon)$. The assumptions on u yield $u \in W^{1,\gamma}(\Omega \setminus \overline{S}_u)$, so that setting $u_\epsilon = u\psi_\epsilon$ we have $u_\epsilon \in W^{1,\gamma}(\Omega)$ and $\sigma_\epsilon \nabla u_\epsilon \in W^{1,p}(\Omega;\mathbf{R}^n)$.

The cut-off function ψ_ϵ is built only in the tubular neighborhood of S_u, otherwise the term $\xi_\epsilon \int (s_\epsilon^2 + \zeta_\epsilon)|\nabla u_\epsilon|^\gamma dx$ cannot be controlled in the neighborhood of $S_{\nabla u} \setminus S_u$. Then we must set $\kappa_\epsilon = 0$ otherwise $\mathcal{F}_\epsilon(w_\epsilon)$ is not finite.

With these choices the estimates proceed in the same way as in the proof of Theorem 3.3 taking into account that the upper limit of the term $\xi_\epsilon \int (s_\epsilon^2+\zeta_\epsilon)|\nabla u_\epsilon|^\gamma dx$ does not exceed

$$\xi_\epsilon \int_\Omega |\nabla u|^\gamma\, dx + \limsup_{\epsilon\to 0^+} \xi_\epsilon\zeta_\epsilon \int_{\Omega\cap\{b_\epsilon/2 \le \tau_1 \le b_\epsilon\}} |\nabla u_\epsilon|^\gamma\, dx\ .$$

In view of the assumption on $\xi_\epsilon\zeta_\epsilon$, we can find an infinitesimal b_ϵ faster than ϵ such that $\xi_\epsilon\zeta_\epsilon = o(b_\epsilon^{\gamma-1})$, for instance, $b_\epsilon = (\epsilon(\xi_\epsilon\zeta_\epsilon)^{\frac{1}{\gamma-1}})^{1/2}$, so that the lim sup above is zero. □

8. An application to the computation of depth from stereo images. The Γ-convergent approximation has been experimented on the problem of computation of depth from a pair of stereo images for the purpose of illustration. In the following Ω denotes the open set $(0,1)\times(0,1)$ of $\mathbf{R}^2$, and $x = (x_1, x_2)$. In the case of parallel camera geometry [22] we choose the expression of the function $\Phi(x,u)$ used in [23, 24, 25]:

$$\Phi(x,u) = \mu\left[L(x_1,x_2) - R(x_1+u,x_2)\right]^2\ ,$$

where u is the disparity function, $\mu > 0$ is a parameter, and R, L are bounded continuous functions corresponding to the right and left image intensities. Depth is

inversely proportional to disparity. The Γ-convergence theorem may be applied if the functions R and L satisfy the conditions of the Remark 3.5, which can be fulfilled, for instance, by means of a convolution of the image intensities with a smooth kernel having a suitably small diameter. For the purpose of illustration we set $\gamma = 2$.

A simple discretization method, commonly used for computer vision problems [35], may be applied to the functionals $\mathcal{F}_\epsilon$ in a straightforward way. Discrete versions of u, s, and σ are defined on a square lattice of coordinates (ih, jh), where $h = 1/(N-1)$, $0 \le i \le N-1$, $0 \le j \le N-1$. We denote by $u^h_{i,j}$, $s^h_{i,j}$, and $\sigma^h_{i,j}$, an approximation of $u(ih, jh)$, $s(ih, jh)$, and $\sigma(ih, jh)$, respectively. We denote by $u^h, s^h, \sigma^h \in \mathbf{R}^{N^2}$ the vectors of the discrete variables. Then we set $\kappa_\epsilon = 0$, $\zeta_\epsilon > 0$, and we discretize

$$\mathcal{F}^1_\epsilon(u, s, \sigma) = \int_\Omega \sigma^2 |\nabla^2 u|^2 \, dx + \xi_\epsilon \int_\Omega (s^2 + \zeta_\epsilon) |\nabla u|^2 \, dx$$

by

$$\begin{aligned} \mathcal{F}^1_{\epsilon,h}(u^h, s^h, \sigma^h) = \sum_{i,j} \Bigg\{ & (\sigma^h_{i,j})^2 \frac{1}{h^2} \left[(u^h_{i+1,j} - 2u^h_{i,j} + u^h_{i-1,j})^2 \right. \\ & + 2(u^h_{i+1,j+1} - u^h_{i,j+1} - u^h_{i+1,j} + u^h_{i,j})^2 \\ & \left. + (u^h_{i,j+1} - 2u^h_{i,j} + u^h_{i,j-1})^2 \right] \\ & + \xi_\epsilon ((s^h_{i,j})^2 + \zeta_\epsilon) \left[(u^h_{i+1,j} - u^h_{i,j})^2 + (u^h_{i,j+1} - u^h_{i,j})^2 \right] \Bigg\}. \end{aligned} \tag{8.1}$$

We set

$$\mathcal{F}^2_\epsilon(s, \sigma) = (\alpha - \beta) \mathcal{G}_\epsilon(s) + \beta \mathcal{G}_\epsilon(\sigma) \ ,$$

and we discretize $\mathcal{G}_\epsilon(s)$ by

$$\mathcal{G}_{\epsilon,h}(s^h) = \sum_{i,j} \left\{ \epsilon \left[(s^h_{i+1,j} - s^h_{i,j})^2 + (s^h_{i,j+1} - s^h_{i,j})^2 \right] + \frac{h^2}{4\epsilon} (s^h_{i,j} - 1)^2 \right\} \ , \tag{8.2}$$

and analogously for $\mathcal{G}_\epsilon(\sigma)$. Then we set

$$\mathcal{F}^3(u) = \mu \int_\Omega \left[L(x_1, x_2) - R(x_1 + u(x_1, x_2), x_2) \right]^2 \, dx \ ,$$

which is discretized by

$$\mathcal{F}^3_h(u^h) = \mu \sum_{i,j} h^2 \left(L^h_{i,j} - R^h_{i+u^h_{i,j},j} \right)^2 \ , \tag{8.3}$$

where $R^h_{i,j}$, $L^h_{i,j}$ denote an approximation of $R(ih, jh)$, $L(ih, jh)$ and, since $u^h_{i,j}$ is generally not an integer, the discretization of R is computed by means of a linear interpolation. We set

$$\mathcal{F}_{\epsilon,h}(u^h, s^h, \sigma^h) = \mathcal{F}^1_{\epsilon,h}(u^h, s^h, \sigma^h) + \mathcal{F}^2_{\epsilon,h}(s^h, \sigma^h) + \mathcal{F}^3_h(u^h) \ .$$

In order to recover a stable solution, the grid must resolve the width of the transition region of the functions s and σ. Then the discretization step should be

at least $h = o(\epsilon)$ as it has been shown in [8] for the discretization of the Ambrosio and Tortorelli approximating functionals. A global solution of the discrete nonconvex variational problem could be computed by means of a stochastic optimization method. However, we use a faster deterministic continuation procedure in which α and β are considered as continuation variables [35]. The functional $\mathcal{F}_\epsilon^1 + \mathcal{F}_\epsilon^2$ becomes increasingly convex for larger values of these variables. Then a solution of the system of equations

$$\nabla \mathcal{F}_{\epsilon,h}(u^h, s^h, \sigma^h) = 0$$

is computed by using a nonlinear Gauss–Seidel iterative method, with α and β initially set to high values, then gradually lowered. The continuation procedure yields experimental good, although not globally optimal, solutions. The parameters α and β are lowered according to the rule

$$\alpha^{(k)} = \alpha_0 (c)^k, \qquad \beta^{(k)} = \beta_0 (c)^k, \tag{8.4}$$

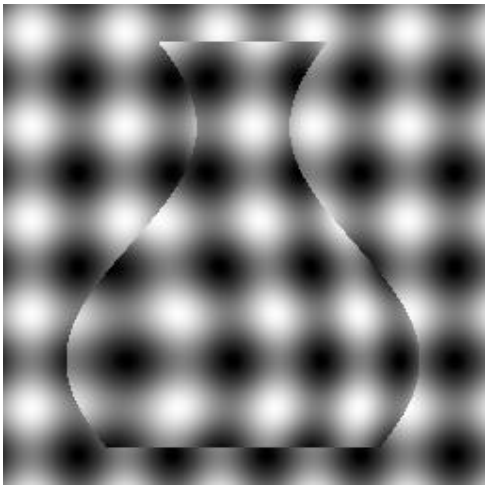
where α_0, β_0 are the initial values, $c < 1$ is a real positive number, and each step k of the continuation procedure consists of 32 nonlinear Gauss–Seidel iterations.

The Γ-convergence theorem and the continuation algorithm have been experimented on synthetic stereo pairs of images corresponding to simple patterns. The images are discretized with N=256. The brightness patterns of all the surfaces represented in the synthetic images are linear combinations of spatially orthogonal sinusoids. The spatial frequency of the sinusoids is chosen to give a reasonably strong brightness gradient such as that usually required for binocular stereo matching (see also March [23, 24, 25]). The range of brightness values for L, R is [0,255]. Depth has to be recovered from the local geometrical distortion of the brightness pattern in the left image relative to the one in the right image. The periodicity of the brightness pattern causes further difficulties to the problem of recovering disparity because of the presence of many ambiguous corresponding points in the two images.

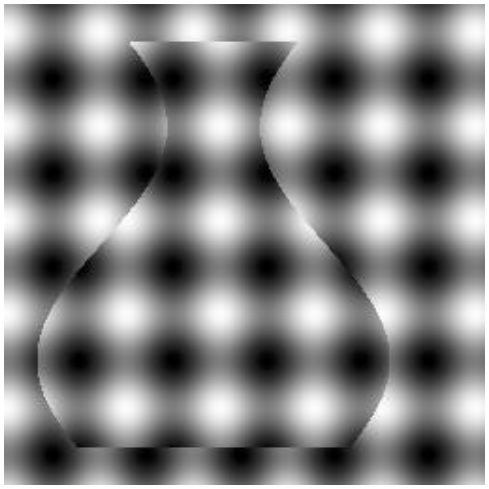
The algorithm was started with an initial estimate of the disparity function u equal to a constant value, and setting the functions σ, s equal to 1 everywhere. The values of the parameters in the functional were chosen on the basis of the results of a number of experiments.

Figures 1(a) and 1(b) show the two images L and R of a stereo pair representing an object shaped as a revolution surface and portrayed against a plane background. The value of disparity ranges from 14 to 32 pixels ($14h \le u \le 32h$). The stereo disparity u in the images of Figures 1(a) and 1(b) is discontinuous along the occluding boundary between the curved surface and the plane background. The function u has no creases in this example.

We set $\sqrt{\mu} = 26$ and $c = 0.8$ in (8.4). The continuation procedure is iterated for 43 steps (1376 total Gauss–Seidel iterations) and the final values of the parameters are $\alpha = 37$, $\beta = 32.7$. Figure 2(a) shows the function σ computed with $\epsilon = 2 \cdot 10^{-2}$, and Figure 2(b) shows the same function computed with $\epsilon = 1.3 \cdot 10^{-2}$. Figure 3(a) shows the function σ computed with $\epsilon = 6.5 \cdot 10^{-3}$: in this case σ reaches values of order 10^{-5} along the discontinuity set of u. In the figures representing the functions σ and s by means of grey values, white corresponds to 0 and black corresponds to 1. The figures show the convergence of the functions σ_ϵ towards the discontinuity set of the disparity u as ϵ decreases, thus illustrating the behavior of Γ-convergence in this specific example. Figure 3(b) shows the function s computed with $\epsilon = 6.5 \cdot 10^{-3}$. Because of the presence of the factor ξ_ϵ converging to zero, the values of the functions s_ϵ might approach zero more slowly than the functions σ_ϵ as ϵ tends to zero.



(a) L image of a synthetic stereo pair (only jumps).



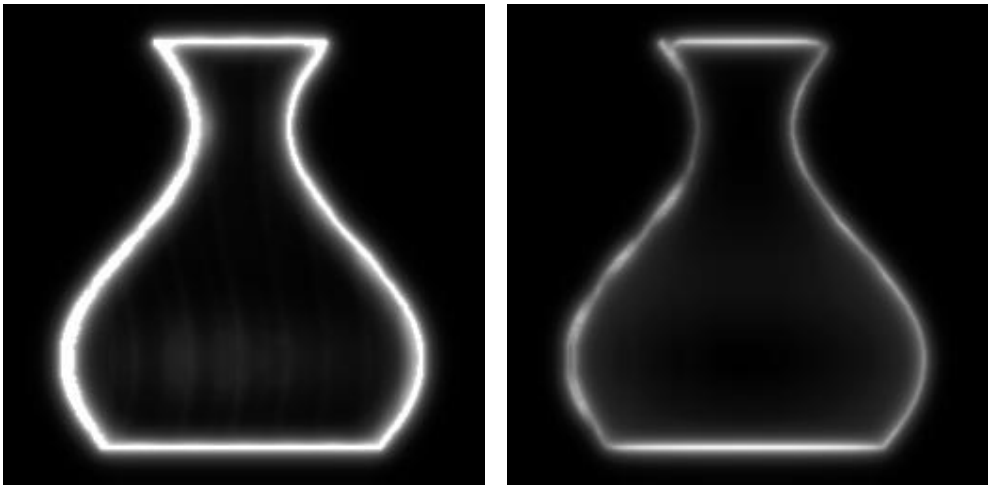
(b) R image of a synthetic stereo pair (only jumps).

FIG. 1.

(a) The function σ computed with $\epsilon = 2.0 \cdot 10^{-2}$.

(b) The function σ computed with $\epsilon = 1.3 \cdot 10^{-2}$.

FIG. 2.



(a) The function σ computed with $\epsilon = 6.5 \cdot 10^{-3}$.

(b) The function s computed with $\epsilon = 6.5 \cdot 10^{-3}$.

FIG. 3.

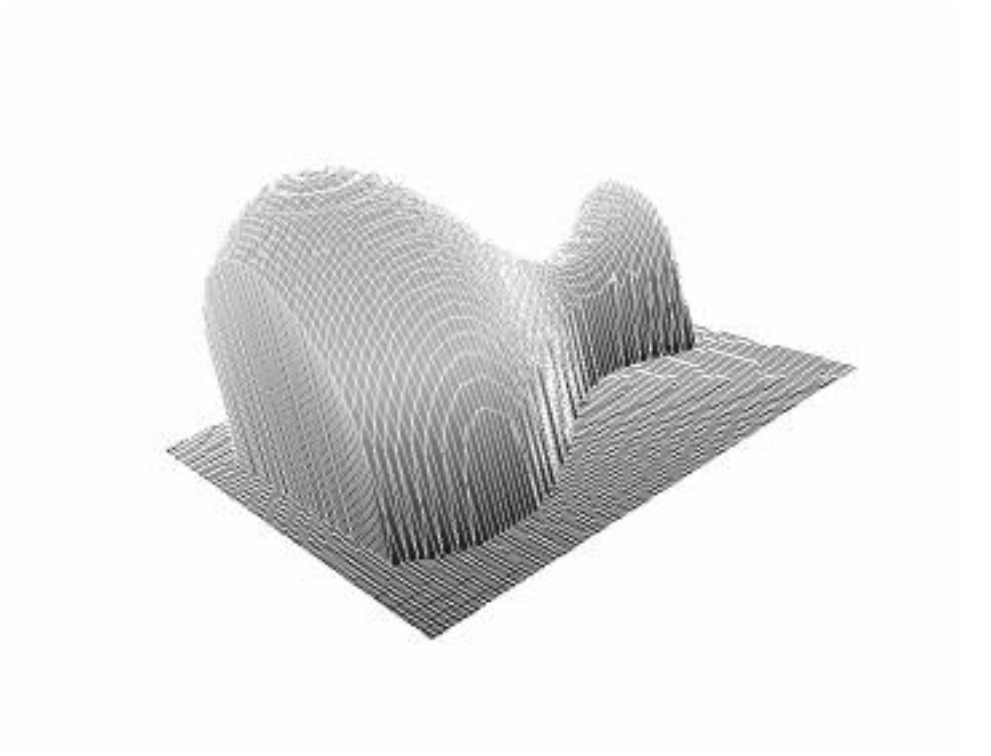
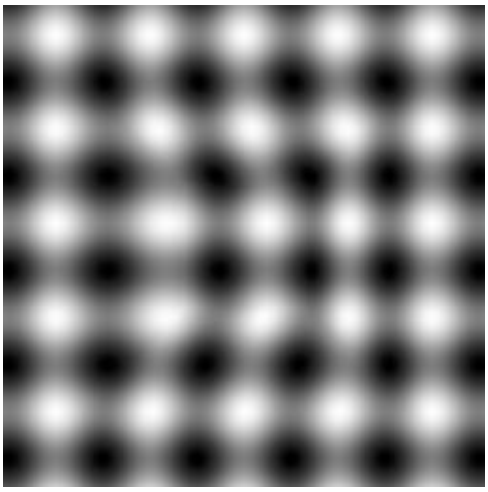
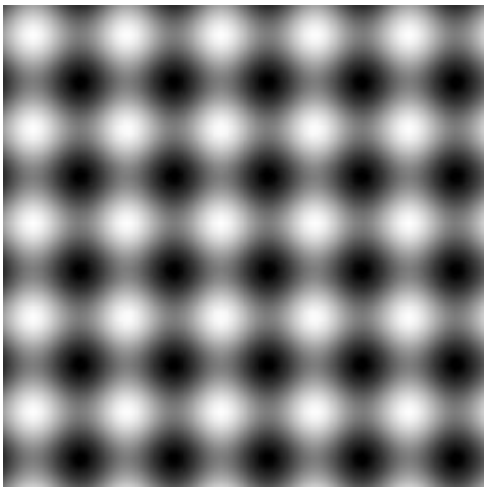


FIG. 4. *Surfaces with jumps recovered from the stereo pair* 1(a), (b).

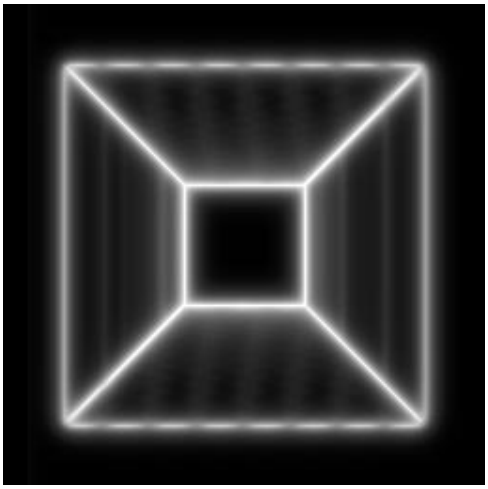


(a) L image of a synthetic stereo pair (only creases).

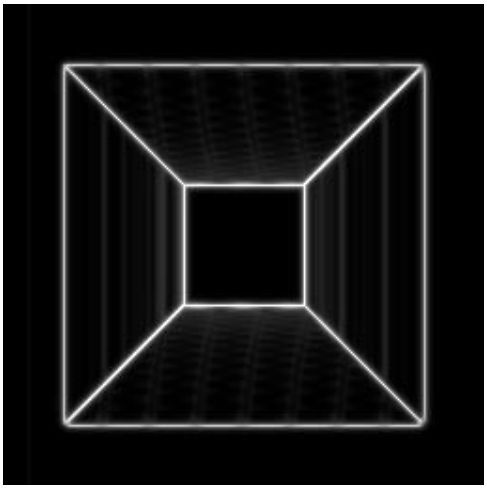


(b) R image of a synthetic stereo pair (only creases).

FIG. 5.



(a) The function σ computed with $\epsilon = 8.0 \cdot 10^{-3}$.



(b) The function σ computed with $\epsilon = 2.0 \cdot 10^{-3}$.

FIG. 6.

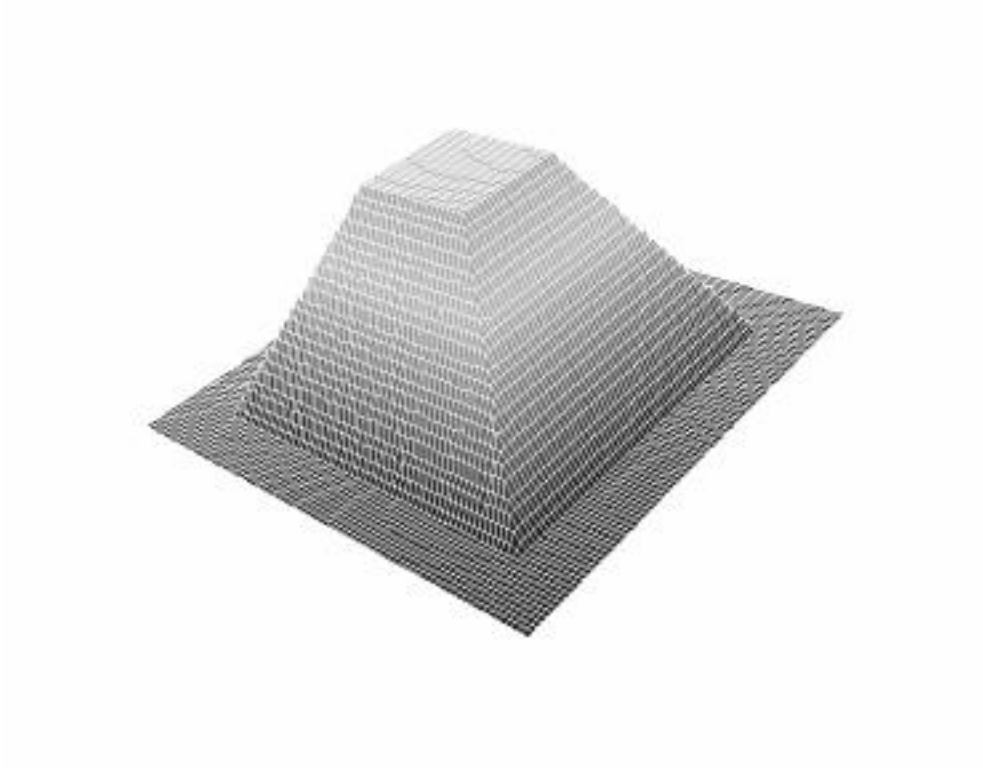
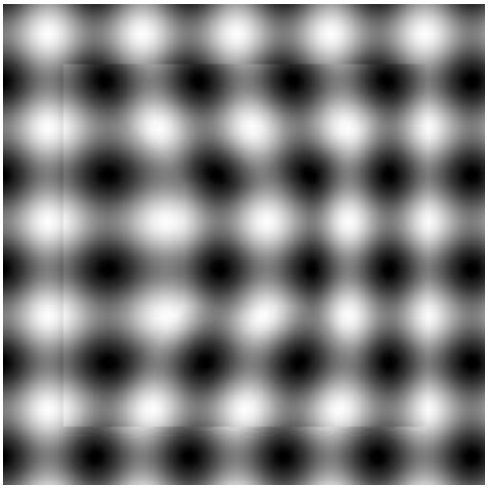
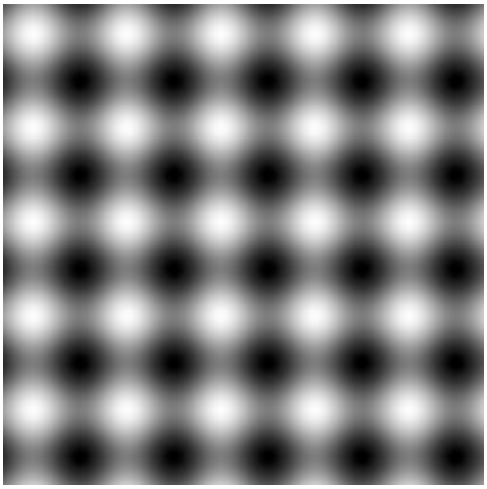


FIG. 7. *Surface with creases recovered from the stereo pair* 5(a), (b).

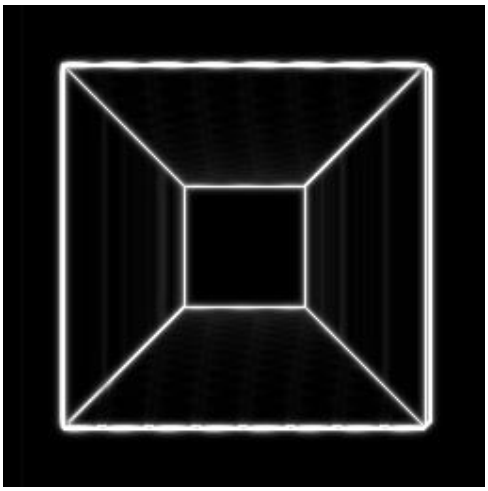


(a) L image of a synthetic stereo pair (jumps+creases).

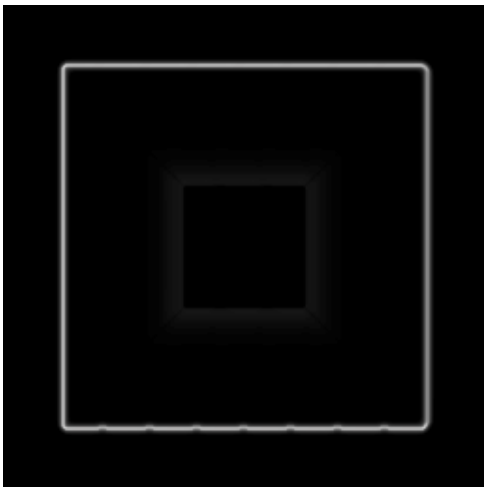


(b) R image of a synthetic stereo pair (jumps+creases).

FIG. 8.



(a) The function σ computed with $\epsilon = 2.0 \cdot 10^{-3}$.



(b) The function s computed with $\epsilon = 2.0 \cdot 10^{-3}$.

FIG. 9.

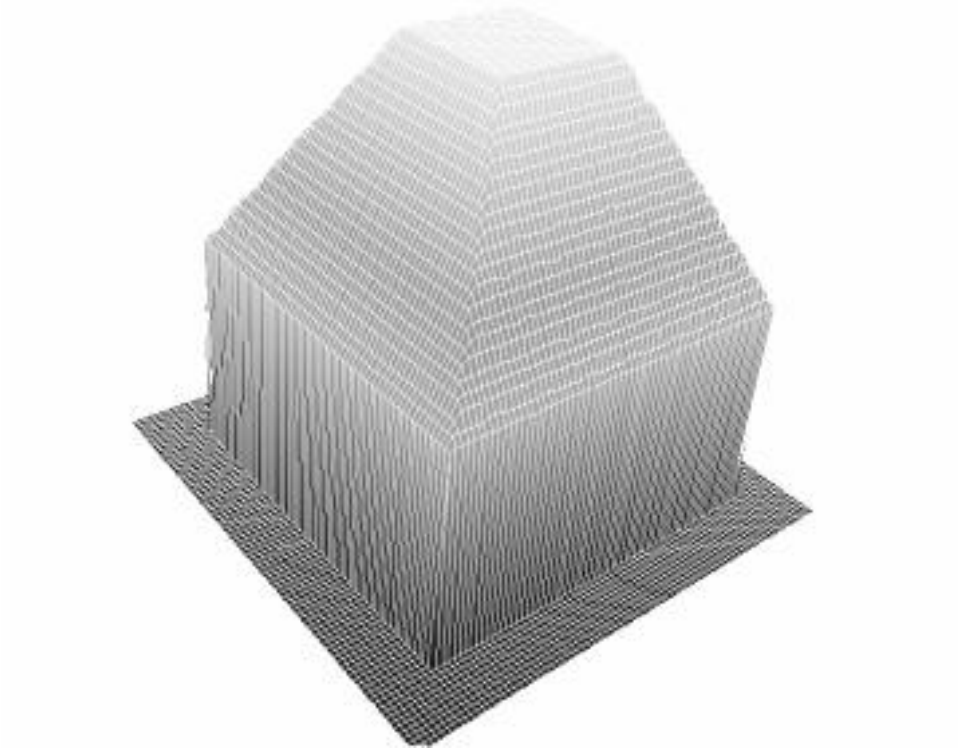


FIG. 10. *Surface with jumps and creases recovered from the stereo pair* 8(a), (b).

For instance, in the present example s reaches the value of $6 \cdot 10^{-2}$, but takes values between 0.3 and 0.7 along some portions of the discontinuity contour. Finally, Figure 4 shows the surfaces corresponding to the disparity map recovered from the stereo pair. The discontinuity set is correctly reconstructed along the occluding contour.

Figures 5(a) and 5(b) show the images L, R of a stereo pair representing a truncated pyramid laid upon a plane background. Disparity ranges from 12 to 24 pixels ($12h \le u \le 24h$) and, in this example, the function u has creases and no jumps.

We set $\sqrt{\mu} = 47$ and $c = 0.76$. The continuation procedure consists of 45 steps (1440 total Gauss–Seidel iterations) and the final values of the parameters are $\alpha = 14.6$, $\beta = 7.3$. Figure 6(a) shows the function σ computed with $\epsilon = 8 \cdot 10^{-3}$. Figure 6(b) shows σ computed with $\epsilon = 2 \cdot 10^{-3}$: in this case σ reaches values of order either 10^{-3} or 10^{-2} along the set of creases of u. The figures show the capability of Γ-convergence in the localization of the creases.

Figure 7 shows the surface recovered from the stereo pair with the creases correctly reconstructed.

Figures 8(a) and 8(b) show the last stereo pair used in the computer experiments which is obtained from the previous one by introducing a jump between the truncated pyramid and the plane background. Disparity ranges from 8 to 24 pixels ($8h \le u \le 24h$). In this example the function u has both creases and jumps.

We set $\sqrt{\mu} = 47$ and $c = 0.76$. The continuation procedure consists of 43 steps (1376 Gauss–Seidel iterations) and the final values of the parameters are $\alpha = 14.3$, $\beta = 12.6$. Figures 9(a) and 9(b) show, respectively, the functions σ and s computed with $\epsilon = 2 \cdot 10^{-3}$. The function σ reaches values of order 10^{-5} along the jumps, and of order 10^{-2} along the creases, while s is about 0.3 along the jumps. Finally, Figure 10 shows the surfaces corresponding to the disparity map recovered from the stereo pair. Both discontinuities and gradient discontinuities are reconstructed.

Acknowledgments. The authors wish to thank an anonymous referee for the careful review of the paper which improved the presentation. The third author wishes to thank Dr. Mario Rosati for many discussions.

REFERENCES

[1] L. Ambrosio, *A compactness theorem for a new class of functions of bounded variation*, Boll. Un. Mat. Ital., 3–B (1989), pp. 857–881.

[2] L. Ambrosio, *Existence theory for a new class of variational problems*, Arch. Rational Mech. Anal., 111 (1990), pp. 291–322.

[3] L. Ambrosio, *The space $SBV(\Omega)$ and free discontinuity problems*, in Variational and Free Boundary Problems, A. Friedman and J. Spruck eds., IMA Vol. Math. Appl. 53, Springer, New York, 1993, pp. 29–45.

[4] L. Ambrosio, N. Fusco, and D. Pallara, *Functions of Bounded Variation and Free Discontinuity Problems*, Oxford Math. Monogr., Oxford University Press, New York, 2000.

[5] L. Ambrosio and V. M. Tortorelli, *Approximation of functionals depending on jumps by elliptic functionals via Γ-convergence*, Comm. Pure Appl. Math., 43 (1990), pp. 999–1036.

[6] L. Ambrosio and V. M. Tortorelli, *On the approximation of free discontinuity problems*, Boll. Un. Mat. Ital. B (7), 6 (1992), pp. 105–123.

[7] G. Bellettini and A. Coscia, *Approximation of a functional depending on jumps and corners*, Boll. Un. Mat. Ital. B (7), 8 (1994), pp. 151–181.

[8] G. Bellettini and A. Coscia, *Discrete approximation of a free discontinuity problem*, Numer. Funct. Anal. Optim., 15 (1994), pp. 201–224.

[9] A. Blake and A. Zisserman, *Visual Reconstruction*, MIT Press, Cambridge, MA, 1987.

[10] G. Buttazzo, *Semicontinuity, Relaxation and Integral Representation in the Calculus of Variations*, Pitman Res. Notes Math. Ser. 207, Longman, Harlow, UK, 1989.

[11] A. P. Calderon and A. Zygmund, *On the differentiability of functions which are of bounded*

variation in Tonelli's sense, Rev. Un. Mat. Argentina, 20 (1960), pp. 102–121.
[12] M. Carriero, A. Leaci, and F. Tomarelli, *A second order model in image segmentation: Blake & Zisserman functional*, in Variational Methods for Discontinuous Structures, R. Serapioni and F. Tomarelli, eds., Birkhäuser, Basel, 1996, pp. 57–72.
[13] M. Carriero, A. Leaci, and F. Tomarelli, *Strong minimizers of Blake & Zisserman functional*, Ann. Scuola Norm. Sup. Pisa Cl. Sci., 25 (1997), pp. 257–285.
[14] M. Carriero, A. Leaci, and F. Tomarelli, *Density estimates and further properties of Blake & Zisserman functional*, in From Convexity to Non Convexity, G. Gilbert, P. D. Panagiotopoulos, and P. Pardalos, eds., Kluwer, Dordrecht, The Netherlands, 2000.
[15] G. Dal Maso, *An Introduction to Γ-Convergence*, Progr. Nonlinear Differential Equations Appl. 8, Birkhäuser, Boston, MA, 1993.
[16] E. De Giorgi, *Nuovi teoremi relativi alle misure* $(r-1)$*–dimensionali in uno spazio a* r *dimensioni*, Ricerche Mat., 4 (1955), pp. 95–113.
[17] E. De Giorgi and L. Ambrosio, *Un nuovo funzionale del calcolo delle variazioni*, Atti Accad. Naz. Lincei Cl. Sci. Fis. Mat. Natur. Rend. Lincei (8) Mat. Appl., 82 (1988), pp. 199–210.
[18] E. De Giorgi, *Free discontinuity problems in calculus of variations*, Frontiers in Pure Mathematics, North-Holland, Amsterdam, 1991, pp. 55–62.
[19] I. Ekeland and R. Temam, *Convex Analysis and Variational Problems*, North-Holland, Amsterdam, 1976.
[20] H. Federer, *Geometric Measure Theory*, Springer–Verlag, Berlin, 1969.
[21] E. Giusti, *Minimal Surfaces and Functions of Bounded Variation*, Birkhäuser, Boston, 1984.
[22] B. K. P. Horn, *Robot Vision*, MIT Press, Cambridge, MA, 1986.
[23] R. March, *Computation of stereo disparity using regularization*, Pattern Recognition Lett., 8 (1988), pp. 181–187.
[24] R. March, *A regularization model for stereo vision with controlled continuity*, Pattern Recognition Lett., 10 (1989), pp. 259–263.
[25] R. March, *Visual reconstruction with discontinuities using variational methods*, Image and Vision Computing, 10 (1992), pp. 30–38.
[26] J. M. Morel and S. Solimini, *Variational Models in Image Segmentation*, Birkhäuser, Basel, 1994.
[27] D. Mumford and J. Shah, *Optimal approximations by piecewise smooth functions and associated variational problems*, Comm. Pure Appl. Math., 42 (1989), pp. 577–685.
[28] P. Nesi, *Variational approach to optical flow estimation managing discontinuities*, Image and Vision Computing, 11 (1993), pp. 419–439.
[29] T. J. Richardson, *Scale Independent Piecewise Smooth Segmentation of Images via Variational Methods*, Ph.D. thesis, Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA, 1990.
[30] T. J. Richardson and S. Mitter, *Approximation, computation, and distortion in the variational formulation*, in Geometry Driven Diffusion in Computer Vision, B. Ter Haar Romeny ed., Kluwer, Dordrecht, The Netherlands, 1994, pp. 169–190.
[31] B. Ter Haar Romeny, ed., *Geometry Driven Diffusion in Computer Vision*, Kluwer, Dordrecht, The Netherlands, 1994.
[32] J. Shah, *Uses of elliptic approximations in computer vision*, in Variational Methods for Discontinuous Structures, R. Serapioni and F. Tomarelli, eds., Birkhäuser, Basel, 1996, pp. 19–34.
[33] J. Shah, H. H. Pien, and J. M. Gauch, *Recovery of surfaces with discontinuities by fusing shading and range data within a variational framework*, IEEE Trans. Image Process., 5 (1996), pp. 1243–1251.
[34] S. Teboul, L. Blanc–Féraud, G. Aubert, and M. Barlaud, *Variational approach for edge-preserving regularization using coupled PDE's*, IEEE Trans. Image Process., 7 (1998), pp. 387–397.
[35] D. Terzopoulos, *The computation of visible-surface representations*, IEEE Trans. Pattern Anal. Machine Intell., 10 (1988), pp. 417–438.
[36] A. I. Vol'pert and S. I. Hudjaev, *Analysis in Classes of Discontinuous Functions and Equations of Mathematical Physics*, Martinus Nijhoff Publishers, Dordrecht, The Netherlands, 1985.

SIAM J. MATH. ANAL.
Vol. 32, No. 6, pp. 1198–1226

HOMOGENIZATION OF THIN STRUCTURES BY TWO-SCALE METHOD WITH RESPECT TO MEASURES*

GUY BOUCHITTÉ† AND ILARIA FRAGALÀ‡

Abstract. To the aim of studying the homogenization of low-dimensional periodic structures, we identify each of them with a periodic positive measure μ on $\mathbb{R}^n$. We introduce a new notion of two-scale convergence for a sequence of functions $v_\varepsilon \in L^p_{\mu_\varepsilon}(\Omega;\mathbb{R}^d)$, where Ω is an open bounded subset of $\mathbb{R}^n$, and the measures μ_ε are the ε-scalings of μ, namely, $\mu_\varepsilon(B) := \varepsilon^n \mu(\varepsilon^{-1}B)$. Enforcing the concept of tangential calculus with respect to measures and related periodic Sobolev spaces, we prove a structure theorem for all the possible two-scale limits reached by the sequences $(u_\varepsilon, \nabla u_\varepsilon)$ when $\{u_\varepsilon\} \subset \mathcal{C}^1_0(\Omega)$ satisfy the boundedness condition $\sup_\varepsilon \int_\Omega |u_\varepsilon|^p + |\nabla u_\varepsilon|^p \, d\mu_\varepsilon < +\infty$ and when the measure μ satisfies suitable connectedness properties. This leads us to deduce the homogenized density of a sequence of energies of the form $\int_\Omega j(\frac{x}{\varepsilon}, \nabla u)\, d\mu_\varepsilon$, where $j(y,z)$ is a convex integrand, periodic in y, and satisfying a p-growth condition. The case of two parameter integrals is also investigated, in particular for what concerns the commutativity of the limit process.

Key words. thin structures, homogenization, two-scale convergence, tangential calculus with respect to periodic measures, connectedness

AMS subject classifications. 35B40, 28A33, 73K20

PII. S0036141000370260

1. Introduction.

For a given periodic Radon measure μ on $\mathbb{R}^n$, we study the asymptotic behavior of a sequence of functionals of the form

$$J_\varepsilon(u) := \int_\Omega j\left(\frac{x}{\varepsilon}, \nabla u\right) d\mu_\varepsilon \ , \qquad u \in \mathcal{C}^1_0(\Omega) \ , \tag{1.1}$$

where Ω is a bounded open subset of $\mathbb{R}^n$, the integrand $j = j(y,z)$ is assumed to be periodic μ-measurable in y and convex with a p-growth condition in z, ε is a positive parameter tending to zero, and μ_ε is the rescaled measure $\mu_\varepsilon(B) := \varepsilon^n \mu\big(\frac{B}{\varepsilon}\big)$.

We think of μ as the Hausdorff measure $\mathcal{H}^k$ on a k-dimensional periodic domain of $\mathbb{R}^n$: when the small parameter ε tends to zero, we recover by this way the classical framework of homogenization on perforated domains (when $k = n$), or on reticulated thin elastic structures (when $k < n$) (see [5] for an introduction to the subject). Actually, for an assigned continuous bounded function f on Ω, the minimum problem

$$\inf\left\{ J_\varepsilon(u) - \int_\Omega f u \, d\mu_\varepsilon \ : \ u \in \mathcal{C}^1_0(\Omega) \right\} \tag{$\mathcal{P}_\varepsilon$}$$

is the variational counterpart of an elliptic equation posed, with boundary conditions, on an ε-periodic domain. For instance, in the case when μ is the Lebesgue measure on a perforated domain, the choice of a quadratic energy density $j(y,z) = \frac{1}{2}\sum_{1\le i,j\le n} a_{i,j}(y) z_i z_j$, where $\mathcal{A}(y) = a_{i,j}(y)$ is a symmetric matrix of periodic and

*Received by the editors March 13, 2000; accepted for publication (in revised form) October 4, 2000; published electronically February 28, 2001. This research was partially supported by the Italian GNAFA/INDAM through the project "Problemi di Monge-Kantorovich e Strutture Geometriche Deboli."

http://www.siam.org/journals/sima/32-6/37026.html

†Département de Mathématiques, Université de Toulon et du Var, BP132, 83957 La Garde, Cedex, France (bouchitte@univ-tln.fr).

‡Dipartimento di Matematica, Politecnico di Milano, Piazza Leonardo da Vinci, 32, 20133 Milano, Italy (fragala@mate.polimi.it).

bounded coefficients, has been intensively studied in the literature (see [14], [25], [9]). More precisely, the classical setting corresponds to the choice $\mu = \mathcal{L}^n \llcorner A$, where A is the complement of a periodic system of holes; denoting by A_ε the set εA, occupied by the material and by T_ε the collection of holes $A \setminus A_\varepsilon$, the minimum problem $(\mathcal{P}_\varepsilon)$ corresponds in this case to the system

$$(\mathcal{S}_\varepsilon) \qquad \begin{cases} -\operatorname{div}\left(\mathcal{A}(\frac{x}{\varepsilon})\nabla u_\varepsilon\right) = f & \text{in } A_\varepsilon, \\ u_\varepsilon = 0 & \text{on } \partial\Omega, \\ \dfrac{\partial u_\varepsilon}{\partial n} = 0 & \text{on } \partial T_\varepsilon. \end{cases}$$

Under suitable geometric assumptions on T_ε , the behavior of $(\mathcal{S}_\varepsilon)$ when $\varepsilon \to 0$ is well known (see, for instance, [16]) and gives rise to a limit problem of the same type, $-\operatorname{div}(\mathcal{A}_{\text{eff}}\nabla u) = f$; the constant matrix $\mathcal{A}_{\text{eff}}$ represents the physical characteristics of a homogenized material which macroscopically behaves like the inhomogeneous perforated body as $\varepsilon \to 0$.

The same is the essence of homogenization in the case of thin periodic structures, except that another parameter δ must be considered, which corresponds to the thickness of the region occupied by the material; thus, a double passage to the limit is needed, as the small parameters ε and δ both tend to zero [17], [3], [15], [18]. For instance, the case of a one-dimensional ε-periodic structure εS, reinforced by some material distributed along bars of cross-section δ, corresponds in our framework to the choice $\mu = \mathcal{H}^1 \llcorner S$, and possibly to $\mu_\delta = |S_\delta|^{-1}\mathcal{L}^n \llcorner S_\delta$, being $S_\delta := \{x \in \mathbb{R}^n : \operatorname{dist}(x, S) \le \delta\}$.

The classical approach to this kind of problem is based on the possibility of defining a suitable extension of the solution u_ε of each elliptic problem such as $(\mathcal{S}_\varepsilon)$ to the whole $\mathbb{R}^n$ [16], [23]; this assumption strongly depends on the shape of the domain under consideration, so that the proof technique must be adapted in each case.

Here we adopt a different point of view, in order to study the Γ-convergence of the functionals J_ε defined in (1.1). This seems to be an unifying approach for structures of arbitrary dimension, as well as for structures containing junctions between parts of different dimensions (see [18]). We represent the structure by general periodic Radon measure μ and we make use of the tangential calculus with respect to μ, which has been developed in some recent papers [7], [6]. In particular, we introduce a notion of two-scale convergence with respect to measures, which generalizes the method proposed in [26], [1], [2]; we prove a structure result for the two-scale possible limits of a sequence $\{(u_\varepsilon, \nabla u_\varepsilon)\}$, when $u_\varepsilon \in \mathcal{C}_0^1(\Omega)$ satisfy the uniform estimate $\int (|u_\varepsilon|^p + |\nabla u_\varepsilon|^p)\, d\mu_\varepsilon \le M$ (see Theorem 4.2). This allows us to deduce the homogenized energy density associated with J_ε in terms of a local unit-cell problem (see Theorem 5.2); further, when fattened stuctures are considered, the limit process as the two parameters ε (of periodicity) and δ (of thickness) tend to zero, can be shown to be commutative (see Theorem 6.1 and Remark 6.2).

Our main assumptions involve some concepts of *connectedness* on the measure μ. They can be formulated by using the corresponding periodic Sobolev spaces $H^{1,p}_\mu(\mathbf{T})$ on the n-torus $\mathbf{T}$, and are related to the connectedness notions for an open domain of $\mathbb{R}^n$ discussed in [28] as an alternative to the extension operator technique.

The case when μ is "nonconnected" is not treated within; neverthless, we point out that in such a situation the two-parameter variational integrals involved in the above described case of fattened thin structures may give rise to a nonlocal limit energy. This phenomenon deserves to our advice further investigation, as it throws some light

on the nonindifference in letting first the periodicity or the thickness parameter tend to zero.

A short outline of the paper follows.

In section 2 we introduce the new notion of two-scale convergence with respect to measures, and some related results are given, which will be useful for the proof of the main homogenization theorem. In section 3 we develop the theory of the periodic Sobolev spaces associated with a measure μ; this allows us, in particular, to formulate in a natural way some different notions of connectedness for μ, which are briefly discussed in section 4 before giving the statement and proof of the main structure theorem for a two-scale limit of a sequence of gradients. The homogenization result for thin structures is then deduced in section 5 as well as a corrector result in the case of quadratic energies. Finally, section 6 is devoted to the case of a sequence of two-parameter variational integrals. We first show that, when μ is connected, the limit energy is independent of the order we let the two parameters ε and δ tend to zero. We conclude by a counterexample in the nonconnected case.

2. Two-scale convergence with respect to a measure. We fix some preliminary notation. Let Ω be an open bounded subset of $\mathbb{R}^n$ with smooth boundary and let Y be the unitary cube of $\mathbb{R}^n$. We will always assume that μ is a positive Y-periodic Radon measure on $\mathbb{R}^n$ with $\mu(\partial Y) = 0$; notice that the latter condition is not restrictive since Y may be replaced by any translated cell $y_0 + Y$. For any $\varepsilon > 0$, we denote by μ_ε the measure defined by

$$\int_\Omega \varphi(x)\, d\mu_\varepsilon(x) = \varepsilon^n \int_{\Omega/\varepsilon} \varphi(\varepsilon x)\, d\mu(x) \qquad \forall \varphi \in \mathcal{C}_0(\Omega), \tag{2.1}$$

$\mathcal{C}_0(\Omega)$ being the space of continuous and compactly supported functions on Ω.

It is easy to check, using (2.1) and the periodicity of μ, that

$$\mu_\varepsilon \rightharpoonup \mu(Y)\, \mathcal{L}^n \llcorner \Omega \tag{2.2}$$

in the vague topology of measures. To have $\mu_\varepsilon \rightharpoonup \mathcal{L}^n \llcorner \Omega$, we assume without loss of generality that $\mu(Y) = 1$. We also set for brevity $m := (\mathcal{L}^n \llcorner \Omega) \otimes (\mu \llcorner Y)$.

We call $\mathbf{T}$ the n-dimensional torus $\mathbb{R}^n/\mathbf{Z}^n$, and throughout the paper we identify functions on $\mathbf{T}$ with Y-periodic functions on $\mathbb{R}^n$. In particular, by writing $\varphi \in \mathcal{D}(\Omega; \mathcal{C}^\infty(\mathbf{T}))$, we mean that $\varphi = \varphi(x, y)$ is smooth in both its variables $(x, y) \in \Omega \times \mathbb{R}^n$, compactly supported in x and Y-periodic in y. We can now set the following.

Definition 2.1. *Let $u_\varepsilon \in L^p_{\mu_\varepsilon}(\Omega)$ and $u_0 \in L^p_m(\Omega \times \mathbf{T})$ for some $p \geq 1$. We say that $\{u_\varepsilon\}$ two-scale converges to u_0 (with respect to μ and as $\varepsilon \to 0$), and we write $u_\varepsilon \rightharpoonup\!\!\!\rightharpoonup u_0$ if*

$$\lim_{\varepsilon \to 0} \int_\Omega u_\varepsilon(x) \varphi\left(x, \frac{x}{\varepsilon}\right) d\mu_\varepsilon(x) = \int_{\Omega \times Y} u_0(x, y) \varphi(x, y)\, dm(x, y) \quad \forall \varphi \in \mathcal{D}(\Omega; \mathcal{C}^\infty(\mathbf{T}))\ .$$

Remark 2.2. Note that, if $u_\varepsilon \rightharpoonup\!\!\!\rightharpoonup u_0$, then $u_\varepsilon \mu_\varepsilon \rightharpoonup u \mathcal{L}^n \llcorner \Omega$, where $u(x) := \int_Y u_0(x, y)\, d\mu(y)$ (choose in Definition 2.1 a test function $\varphi(x, y) = \psi(x) \in \mathcal{D}(\Omega)$).

On the other hand, similarly as in the well-known case where $\mu_\varepsilon = \mathcal{L}^n \llcorner \Omega$ for every ε, Definition 2.1 retains more information on the sequence $\{u_\varepsilon \mu_\varepsilon\}$ than the pure weak convergence in the sense of measures: it also takes into account the "oscillations" which have the same frequency as the test functions $\varphi(x, \frac{x}{\varepsilon})$. For instance, in the simplest case when $u_\varepsilon(x) = u(x, \frac{x}{\varepsilon})$, with u continuous on $\Omega \times \mathbb{R}^n$ and Y-periodic in

its second variable, we have $u_\varepsilon \rightharpoonup u$. This follows from the convergence

$$v\left(x,\frac{x}{\varepsilon}\right)\mu_\varepsilon \rightharpoonup \left(\int_Y v(x,y)\,d\mu(y)\right)\mathcal{L}^n \llcorner \Omega\ , \tag{2.3}$$

holding for any function v continuous on $\Omega\times\mathbb{R}^n$ and Y-periodic in y.

In view of the following compactness property, the two-scale convergence is a good notion for a sequence $\{u_\varepsilon\}$ which satisfies a uniform bound of the type $\int_\Omega |u_\varepsilon|^p\,d\mu_\varepsilon \le M$, with $p>1$. Throughout this section, the symbols $\|\cdot\|_{p,\mu_\varepsilon}$ and $\|\cdot\|_{p,m}$ will be used, respectively, for the $L^p_{\mu_\varepsilon}(\Omega)$ and the $L^p_m(\Omega\times Y)$ norms, while $p'=\frac{p}{p-1}$ will denote the conjugate exponent of p.

PROPOSITION 2.3. *Let $p>1$, and let $u_\varepsilon\in L^p_{\mu_\varepsilon}(\Omega)$ satisfy $\int_\Omega |u_\varepsilon|^p\,d\mu_\varepsilon\le M$. Then there exists a subsequence u_{ε_k} such that $u_{\varepsilon_k}\rightharpoonup u_0$, with $u_0\in L^p_m(\Omega\times Y)$; in particular, $u_{\varepsilon_k}\mu_{\varepsilon_k}\rightharpoonup u\mathcal{L}^n$, where $u(x):=\int_Y u_0(x,y)\,d\mu(y)$.*

Proof. Let $(B\,,\,\|\cdot\|_B)$ be the Banach space of all continuous functions on $\Omega\times Y$ vanishing at the boundary of $\Omega\times Y$, endowed with the uniform norm. For any $\varphi\in B$ we have

$$\|\varphi\left(x,\frac{x}{\varepsilon}\right)\|_{p',\mu_\varepsilon}\ \le\ C\|\varphi\|_B\ , \tag{2.4}$$

$$\lim_{\varepsilon\to 0}\ \|\varphi\left(x,\frac{x}{\varepsilon}\right)\|_{p',\mu_\varepsilon}\ =\ \|\varphi(x,y)\|_{p',m}. \tag{2.5}$$

The first inequality follows from the boundedness of μ_ε (as Ω is relatively compact with $\mathcal{L}^n(\partial\Omega)=0$, we have by (2.2) $\mu_\varepsilon(\Omega)\to\mathcal{L}^n(\Omega)$ as $\varepsilon\to 0$). To prove (2.5), we simply apply (2.3) with $v(x,y)=|\varphi(x,y)|^{p'}$. Now let us consider for any $\varepsilon>0$ the linear operator T_ε defined on B by

$$\langle T_\varepsilon,\varphi\rangle := \int_\Omega u_\varepsilon(x)\varphi\left(x,\frac{x}{\varepsilon}\right)\,d\mu_\varepsilon,$$

where we think of φ as implicitly extended by Y-periodicity on $\Omega\times\mathbb{R}^n$. By the Hölder inequality, the assumption on u_ε and (2.4), we have

$$\begin{aligned}|\langle T_\varepsilon,\varphi\rangle| \le \|u_\varepsilon\|_{p,\mu_\varepsilon}\|\varphi\left(x,\frac{x}{\varepsilon}\right)\|_{p',\mu_\varepsilon} &\le M^{1/p}\|\varphi\left(x,\frac{x}{\varepsilon}\right)\|_{p',\mu_\varepsilon}\\ &\le M^{1/p}C\|\varphi\|_B.\end{aligned} \tag{2.6}$$

Thus T_ε is a bounded sequence in the space B' of finite Borel measures and, possibly passing to a subsequence, we can assume that T_ε weakly star converges to an element T_0 of B'. Then, using (2.5) to pass to the limit as $\varepsilon\to 0$ in (2.6), we get

$$|\langle T_0,\varphi\rangle|\le M^{1/p}\|\varphi(x,y)\|_{p',m}\qquad\forall\varphi\in B.$$

Hence, by the density of B in $L^{p'}_m(\Omega\times Y)$, T_0 may be represented by an element $u_0\in L^p_m(\Omega\times Y)$ satisfying

$$\lim_{\varepsilon\to 0}\int_\Omega u_\varepsilon(x)\varphi\left(x,\frac{x}{\varepsilon}\right)\,d\mu_\varepsilon=\int_{\Omega\times Y}u_0(x,y)\varphi(x,y)\,dm\qquad\forall\varphi\in B.$$

This also implies, by Remark 2.2, the second part of the statement. ∎

Remark 2.4. Under the assumptions of Proposition 2.3, the space of admissible test functions in the two-scale convergence can be considerably enlarged. Indeed,

the same proof of Proposition 2.3 works whenever the space of continuous functions vanishing at the boundary of $\Omega \times Y$ is replaced by another separable Banach space B, dense in $L^{p'}_m(\Omega \times Y)$, such that (2.4) and (2.5) hold. For instance, all these conditions are satisfied if we take $B = \mathcal{C}(\overline{\Omega}, L^p_\mu(\mathbf{T}))$ (a detailed proof can be derived with minor changes from [1, Corollary 5.4]); in particular, the convergence (2.3) holds for any $v \in \mathcal{C}(\overline{\Omega}, L^p_\mu(\mathbf{T}))$.

The two-scale convergence enjoys the following lower semicontinuity property when dealing with convex periodic integrands.

PROPOSITION 2.5. *Let $v_\varepsilon \in L^p_{\mu_\varepsilon}(\Omega; \mathbb{R}^d)$ two-scale converge (by components) to $v_0 \in L^p_m(\Omega \times Y; \mathbb{R}^d)$, $p > 1$; let $j = j(y,z)$ be a function on $\mathbb{R}^n \times \mathbb{R}^d$ μ-measurable and Y-periodic in y, convex in z, and satisfying, for some positive constants C, c, the estimate*

$$c|z|^p \le j(y,z) \le C\,(1 + |z|^p) \qquad \forall (y,z) \in \mathbb{R}^n \times \mathbb{R}^d\ .$$

Then

$$\liminf_{\varepsilon \to 0} \int_\Omega j\left(\frac{x}{\varepsilon}, v_\varepsilon\right) d\mu_\varepsilon \ge \int_{\Omega \times Y} j(y, v_0(x,y))\, dm.$$

Proof. For any function $\varphi \in \mathcal{D}(\Omega \times Y)$, the Fenchel's inequality gives

$$\int_\Omega j\left(\frac{x}{\varepsilon}, v_\varepsilon\right) d\mu_\varepsilon \ge \int_\Omega v_\varepsilon(x) \varphi\left(x, \frac{x}{\varepsilon}\right) d\mu_\varepsilon - \int_\Omega j^*\left(\frac{x}{\varepsilon}, \varphi\left(x, \frac{x}{\varepsilon}\right)\right) d\mu_\varepsilon,$$

where $j^*(y, \cdot)$ denotes the conjugate functional of $j(y, \cdot)$, and φ is extended by Y-periodicity on $\Omega \times \mathbb{R}^n$. Since $v_\varepsilon \rightharpoonup\!\!\!\!\rightharpoonup v_0$, recalling Remark 2.4, passing to the limit as $\varepsilon \to 0$ in the above inequality yields

$$\liminf_{\varepsilon \to 0} \int_\Omega j\left(\frac{x}{\varepsilon}, v_\varepsilon\right) d\mu_\varepsilon \ge \int_{\Omega \times Y} v_0(x,y) \varphi\big(x,y\big)\, dm - \int_{\Omega \times Y} j^*\Big(y, \varphi\big(x,y\big)\Big)\, dm. \tag{2.7}$$

Taking the supremum of the right-hand side of (2.7) when φ varies in $\mathcal{D}(\Omega \times Y)$, one gets, using a classical localization argument and the convexity assumption on j,

$$\liminf_{\varepsilon \to 0} \int_\Omega j\left(\frac{x}{\varepsilon}, v_\varepsilon\right) d\mu_\varepsilon \ \ge \ \int_{\Omega \times Y} j^{**}(y, v_0(x,y))\, dm \ = \int_{\Omega \times Y} j(y, v_0(x,y))\, dm. \qquad \square$$

Applying the above proposition with $j(y,z) = |z|^p$, we get

$$\liminf_{\varepsilon \to 0} \int_\Omega |u_\varepsilon|^p\, d\mu_\varepsilon \ge \int_{\Omega \times Y} |u_0|^p\, dm, \tag{2.8}$$

whenever $u_\varepsilon \in L^p_{\mu_\varepsilon}(\Omega)$ two-scale converge to $u_0 \in L^p(\Omega \times Y)$ $(p > 1)$.

DEFINITION 2.6. *Let $u_\varepsilon \in L^p_{\mu_\varepsilon}(\Omega)$ and $u_0 \in L^p_m(\Omega \times \mathbf{T})$ for some $p > 1$. We say that $\{u_\varepsilon\}$ two-scale strongly converges to u_0 (with respect to μ and as $\varepsilon \to 0$), and we write $u_\varepsilon \rightrightarrows u_0$ if*

$$u_\varepsilon \rightharpoonup\!\!\!\!\rightharpoonup u_0 \quad \text{and} \quad \limsup_{\varepsilon \to 0} \int_\Omega |u_\varepsilon|^p\, d\mu_\varepsilon \le \int_{\Omega \times Y} |u_0|^p\, dm\ .$$

In view of (2.8), when $u_\varepsilon \rightarrow\!\!\!\!\rightarrow u_0$, the $L^p_{\mu_\varepsilon}(\Omega)$-norm of u_ε converges to the $L^p_m(\Omega \times Y)$-norm of u_0. This means that all the oscillations of the sequence $\{u_\varepsilon \mu_\varepsilon\}$ are captured by the two-scale limit, namely, they are in resonance with those of the test functions $\varphi(x, \frac{x}{\varepsilon})$.

Example 2.7. Let $u_\varepsilon = u_0(x, \frac{x}{\varepsilon})$, where $u_0(x, y)$ is given by $u_0(x, y) = u(x)w(y)$, with $u \in L^p(\Omega)$ and $w \in L^\infty(\mathbf{T})$. Then it is easy to check that $u_\varepsilon \rightarrow\!\!\!\!\rightarrow u_0$. In fact this conclusion also holds for more general u_0, in particular, for $u_0 \in C(\overline{\Omega}; L^p_\mu(\mathbf{T}))$.

PROPOSITION 2.8. *Let $p > 1$, and let $u_\varepsilon \in L^p_{\mu_\varepsilon}(\Omega)$, $v_\varepsilon \in L^{p'}_{\mu_\varepsilon}(\Omega)$ satisfy*

$$u_\varepsilon \rightarrow\!\!\!\!\rightarrow u_0 \ , \quad u_0 \in L^p_m(\Omega \times Y) \ , \tag{2.9}$$

$$v_\varepsilon \rightharpoonup\!\!\!\!\rightharpoonup v_0 \ , \quad v_0 \in L^{p'}_m(\Omega \times Y) \ , \quad \int_\Omega |v_\varepsilon|^{p'} \, d\mu_\varepsilon \leq M \ . \tag{2.10}$$

Then

$$u_\varepsilon v_\varepsilon \mu_\varepsilon \rightharpoonup \left(\int_Y u_0(\cdot, y) v_0(\cdot, y) \, d\mu(y) \right) \mathcal{L}^n \llcorner \Omega, \tag{2.11}$$

$$\lim_{\varepsilon \to 0} \int_\Omega |u_\varepsilon - u_0\big(x, \frac{x}{\varepsilon}\big)|^p \, d\mu_\varepsilon = 0 \quad \textit{whenever } u_0 \in \mathcal{C}(\Omega \times \mathbf{T}). \tag{2.12}$$

Proof. Let $\{\varphi_h\} \subset \mathcal{D}(\Omega \times Y)$ be a sequence of smooth functions converging to u_0 in $L^p_m(\Omega \times Y)$ and extended by Y-periodicity on $\Omega \times \mathbb{R}^n$. For any test function $\psi \in C_0(\Omega)$, it holds that

$$\lim_{\varepsilon \to 0} \int_\Omega u_\varepsilon v_\varepsilon \psi \, d\mu_\varepsilon = \lim_{h \to +\infty} \lim_{\varepsilon \to 0} \left\{ \int_\Omega \left[u_\varepsilon - \varphi_h \left(x, \frac{x}{\varepsilon} \right) \right] v_\varepsilon \psi \, d\mu_\varepsilon + \int_\Omega \varphi_h \left(x, \frac{x}{\varepsilon} \right) v_\varepsilon \psi \, d\mu_\varepsilon \right\}.$$

By (2.10) and the choice of $\{\varphi_h\}$ we have

$$\begin{aligned} \lim_{h \to +\infty} \lim_{\varepsilon \to 0} \int_\Omega \varphi_h \left(x, \frac{x}{\varepsilon} \right) v_\varepsilon \psi \, d\mu_\varepsilon &= \lim_{h \to +\infty} \int_{\Omega \times Y} \varphi_h\big(x, y\big) v_0(x, y) \psi(x) \, dm \\ &= \int_{\Omega \times Y} u_0(x, y) v_0(x, y) \psi(x) \, dm. \end{aligned}$$

To prove (2.11), it is then enough to show that

$$\lim_{h \to +\infty} \lim_{\varepsilon \to 0} \int_\Omega \left[u_\varepsilon - \varphi_h \left(x, \frac{x}{\varepsilon} \right) \right] v_\varepsilon \psi \, d\mu_\varepsilon = 0. \tag{2.13}$$

Applying the Hölder inequality and assumption (2.10), we deduce

$$\left| \int_\Omega \left[u_\varepsilon - \varphi_h \left(x, \frac{x}{\varepsilon} \right) \right] v_\varepsilon \psi \, d\mu_\varepsilon \right| \leq M^{1/p'} \|\psi\|_\infty \left\| u_\varepsilon - \varphi_h \left(x, \frac{x}{\varepsilon} \right) \right\|_{p, \mu_\varepsilon}.$$

The Clarkson inequalities give, respectively, for $p \geq 2$ and for $p \leq 2$

$$\begin{cases} \left\| u_\varepsilon - \varphi_h\left(x, \dfrac{x}{\varepsilon}\right)\right\|^p_{p,\mu_\varepsilon} \\ \leq\ 2^p \left\{ \dfrac{1}{2}\|u_\varepsilon\|^p_{p,\mu_\varepsilon} + \dfrac{1}{2}\left\|\varphi_h\left(x, \dfrac{x}{\varepsilon}\right)\right\|^p_{p,\mu_\varepsilon} - \left\| \dfrac{u_\varepsilon + \varphi_h\big(x, \frac{x}{\varepsilon}\big)}{2}\right\|^p_{p,\mu_\varepsilon} \right\}, \\ \left\| u_\varepsilon - \varphi_h\left(x, \dfrac{x}{\varepsilon}\right)\right\|^{p'}_{p,\mu_\varepsilon} \\ \qquad \leq\ 2^{p-1} \left\{ \left[\dfrac{1}{2}\|u_\varepsilon\|^p_{p,\mu_\varepsilon} + \dfrac{1}{2}\left\|\varphi_h\left(x, \dfrac{x}{\varepsilon}\right)\right\|^p_{p,\mu_\varepsilon}\right]^{\frac{1}{p-1}} - \left\| \dfrac{u_\varepsilon + \varphi_h\big(x, \frac{x}{\varepsilon}\big)}{2}\right\|^{p'}_{p,\mu_\varepsilon} \right\}. \end{cases}$$

Using (2.9) and the smoothness of each φ_h, we infer

$$\begin{cases} \limsup\limits_{\varepsilon\to 0} \|u_\varepsilon - \varphi_h\left(x, \dfrac{x}{\varepsilon}\right)\|^p_{p,\mu_\varepsilon} \leq 2^p \left\{ \dfrac{1}{2}\|u_0\|^p_{p,m} + \dfrac{1}{2}\|\varphi_h\|^p_{p,m} - \left\| \dfrac{u_0 + \varphi_h}{2}\right\|^p_{p,m} \right\}, \\ \limsup\limits_{\varepsilon\to 0} \|u_0 - \varphi_h\|^{p'}_{p,m} \leq 2^{p-1} \left\{ \left[\dfrac{1}{2}\|u_0\|^p_{p,m} + \dfrac{1}{2}\|\varphi_h\|^p_{p,m}\right]^{\frac{1}{p-1}} - \left\| \dfrac{u_0 + \varphi_h}{2}\right\|^{p'}_{p,m} \right\}. \end{cases}$$

Passing now to the limsup as $h \to +\infty$, by the choice of the sequence $\{\varphi_h\}$, both the right-hand sides of the above inequalities tend to zero, so that (2.13) is proved.

To obtain (2.12), it is enough to write the Clarkson inequalities with φ_h replaced by u_0 and to observe that, being u_0 globally continuous and Y-periodic in its second variable, the right-hand sides converge to zero as $\varepsilon \to 0$. □

3. The periodic Sobolev spaces $\boldsymbol{H^{1,p}_\mu(\mathbf{T})}$. First, let us give a brief recall about the Sobolev spaces $H^{1,p}_\mu$ introduced in [7].

Let $\mathcal{D} := \mathcal{C}^\infty_0(\mathbb{R}^n)$ and $\mathcal{D}'$ be the space of distributions on $\mathbb{R}^n$. Let $p\,,\, p' \in [1, +\infty]$ be fixed conjugate exponents. For any $q \geq 1$, let $L^q_\mu := L^q_\mu(\mathbb{R}^n)$, $(L^q_\mu)^n := L^q_\mu(\mathbb{R}^n; \mathbb{R}^n)$, and similarly for $L^q_{\mu,\mathrm{loc}}$, $(L^q_{\mu,\mathrm{loc}})^n$. The notation $\|\cdot\|_{q,\mu}$ will be used throughout the section for the usual norms of L^q_μ and $(L^q_\mu)^n$, as well as for $L^q_\mu(\mathbf{T})$ and $(L^q_\mu(\mathbf{T}))^n$ (implicitly assuming in this case, whenever unambiguous, that the integrals are made over Y).

We recall that, for any $\sigma \in (L^1_{\mu,\mathrm{loc}})^n$, $\mathrm{div}(\sigma\mu) \in \mathcal{D}'$ is defined by

$$\langle \mathrm{div}(\sigma\mu), \psi\rangle_{(\mathcal{D}',\mathcal{D})} := -\int_{\mathbb{R}^n} \sigma \cdot \nabla\psi \, d\mu \qquad \forall \psi \in \mathcal{D}\ .$$

Whenever $\mathrm{div}(\sigma\mu)$ is a measure absolutely continuous with respect to μ with a density belonging to $L^{p'}_\mu$, we write for brevity $\mathrm{div}(\sigma\mu) \in L^{p'}_\mu$, and we denote by $\mathrm{div}_\mu\, \sigma$ the derivative $\frac{d}{d\mu}\,\mathrm{div}(\sigma\mu)$.

It will be useful in the following to notice that there holds

$$\mathrm{div}(\psi\sigma\mu) = \psi\,\mathrm{div}(\sigma\mu) + (\sigma\cdot\nabla\psi)\mu \qquad \forall \psi \in \mathcal{D},\ \forall \sigma \in (L^1_{\mu,\mathrm{loc}})^n\ . \tag{3.1}$$

As in [7], we introduce the class $X^{p'}_\mu$ of all vector fields Φ which are "tangent to μ," given by

$$X^{p'}_\mu := \left\{ \Phi \in (L^{p'}_\mu)^n \ :\ \mathrm{div}(\Phi\mu) \in L^{p'}_\mu \right\}\ ,$$

and $T_\mu^p(x)$ the tangent space of μ at x, obtained as

$$T_\mu^p(x) := \mu - \operatorname{ess} \bigcup \left\{ \Phi(x) \ : \ \Phi \in X_\mu^{p'} \right\} , \qquad x \in \mathbb{R}^n \ .$$

For details on the meaning of the μ-essential union and on the properties of the multifunction T_μ^p, we refer to [7], [21]; in particular, at least for usual measures, $T_\mu^p(x)$ does not depend on the exponent p, and it will be denoted simply by $T_\mu(x)$.

The Sobolev space $H_\mu^{1,p} = H_\mu^{1,p}(\mathbb{R}^n)$ is defined as the completion of $\mathcal{D}$ with respect to the norm $\|\psi\|_{1,p,\mu} := \|\psi\|_{p,\mu} + \|\nabla_\mu \psi\|_{p,\mu}$, where $\nabla_\mu \psi(x) := P_\mu(x)[\nabla \psi(x)]$ is the orthogonal projection $P_\mu(x)$ of $\nabla \psi(x)$ onto $T_\mu(x)$. In other words, it turns out that the linear operator $A : D(A) \subset L_\mu^p \to (L_\mu^p)^n$ defined by $D(A) = \mathcal{D}$ and $A\psi = \nabla_\mu \psi$ is closable [7, Proposition 2.1], and the domain of its unique closed extension $\overline{A}$ is called $H_\mu^{1,p}$. By definition $H_\mu^{1,p}$ is a Banach space, and it is reflexive when $p \in (1, +\infty)$. For any $u \in H_\mu^{1,p}$, we set $\nabla_\mu u := \overline{A} u$. Then a straightforward density argument gives the following useful integration by parts:

$$\int_{\mathbb{R}^n} \nabla_\mu u \cdot \Phi \, d\mu = - \int_{\mathbb{R}^n} u \, \operatorname{div}_\mu \Phi \, d\mu \qquad \forall u \in H_\mu^{1,p}, \ \forall \Phi \in X_\mu^{p'} \ .$$

Notice the adjoint operator $A^* : (L_\mu^{p'})^n \to L_\mu^{p'}$ equals $-\operatorname{div}(P_\mu \sigma \mu)$ when applied to any vector field σ belonging to its domain $D(A^*)$, given by

$$D(A^*) = \left\{ \sigma \in (L_\mu^{p'})^n \ : \ \operatorname{div}(P_\mu \sigma \mu) \in L_\mu^{p'} \right\} = \left\{ \sigma \in (L_\mu^{p'})^n \ : \ P_\mu \sigma \in X_\mu^{p'} \right\} =: Y_\mu^{p'} \ .$$

In particular, when $p > 1$, the following characterization by duality of $H_\mu^{1,p}$ can be deduced from Lemma 3.1 below:

$$H_\mu^{1,p} = \left\{ u \in L_\mu^p \text{ such that } \exists C > 0 \ : \ |\langle u, \operatorname{div}(P_\mu \sigma \mu) \rangle| \le C \|\sigma\|_{p',\mu} \quad \forall \sigma \in Y_\mu^{p'} \right\} .$$

LEMMA 3.1. *Let V, W be two Banach spaces, with W reflexive. Assume that $T : D(T) \subset V \to W$ is a linear and closable operator with dense domain. Then*

$$D(\overline{T}) = \left\{ v \in V \text{ such that } \exists C > 0 \ : \ |\langle v, T^* w' \rangle_{(V,V')}| \le C \|w'\|_{W'} \ \forall w' \in D(T^*) \right\}. \tag{3.2}$$

The proof of this lemma is an immediate consequence of the equality $\overline{T} = T^{**}$, which holds by standard arguments of functional analysis [10].

Let us turn now our attention to Y-periodic functions. Set

$$X_\mu^{p'}(\mathbf{T}) := \left\{ \Phi \in \left(L_\mu^{p'}(\mathbf{T}) \right)^n \ : \ \operatorname{div}(\Phi \mu) \in L_{\mu,\mathrm{loc}}^{p'} \right\} .$$

Notice that, since $\mu(\partial Y) = 0$, we have

$$\mu - \operatorname{ess} \bigcup \left\{ \Phi(x) \ : \ \Phi \in X_\mu^{p'}(\mathbf{T}) \right\} = T_\mu(x) \qquad \mu\text{-a.e. (almost everywhere) on } Y. \tag{3.3}$$

Let $A_\sharp : D(A_\sharp) \subset L_\mu^p(\mathbf{T}) \to (L_\mu^p(\mathbf{T}))^n$ be the linear operator given by $D(A_\sharp) = \mathcal{C}^\infty(\mathbf{T})$, $A_\sharp \psi = \nabla_\mu \psi$. In order to prove that $A_\sharp$ is closable, we need some preliminary lemmas. For any $\psi \in \mathcal{D}$, we set

$$\psi^\sharp(y) = \sum_{i \in \mathbf{Z}^n} \psi(y - i) \ , \qquad y \in Y.$$

The function $\psi^\sharp$ is well defined as soon as ψ is compactly supported, and it turns out to be Y-periodic by construction.

LEMMA 3.2. *For any $v \in L^1_\mu(\mathbf{T})$ and any $\psi \in \mathcal{D}$, there holds $\int_{\mathbb{R}^n} v\psi \, d\mu = \int_Y v\psi^\sharp \, d\mu$.*

Proof. Due to the Y-periodicity of v and μ, we have

$$\int_{\mathbb{R}^n} v\psi \, d\mu = \sum_{i\in\mathbf{Z}^n} \int_{i+Y} v\psi \, d\mu = \sum_{i\in\mathbf{Z}^n} \int_Y v(y)\psi(y-i)\, d\mu(y) \ .$$

The last sum over $i \in \mathbf{Z}^n$ may equivalently be made over the finite set of indices $I_\psi := \{i \in \mathbf{Z}^n \ : \ (Y-i) \cap \operatorname{spt}\psi \neq \emptyset\}$. Thus, it can be passed under the sign of integral and the lemma is proved. □

LEMMA 3.3. *There exists a sequence $\{\psi_h\} \subset \mathcal{D}$ such that $\psi_h^\sharp \to 1$ in $H^{1,\infty}_\mu(\mathbf{T})$.*

Proof. For any $h \in \mathbb{N}$, take $\psi_h \in \mathcal{C}^\infty_0(\mathbb{R}^n; [0, h^{-n}])$ satisfying

$$\psi_h = h^{-n} \ \text{ on } hY \ , \qquad \psi_h = 0 \ \text{ on } \mathbb{R}^n \setminus (h+1)Y \ ; \tag{3.4}$$

$$|\nabla\psi_h| \le Ch^{-n} \ \text{ on } \mathbb{R}^n. \tag{3.5}$$

For any $y \in Y$, we have

$$h^{-n}\operatorname{card} I_h(y) \le \psi_h^\sharp(y) \le h^{-n}\operatorname{card} J_h(y) \ , \tag{3.6}$$

where $I_h(y) = \{i \in \mathbf{Z}^n \ : \ \psi_h(y-i) = h^{-n}\}$, $J_h(y) = \{i \in \mathbf{Z}^n \ : \ \psi_h(y-i) \neq 0\}$.

By (3.4), $\operatorname{card} I_h(y) \ge h^n$ and $\operatorname{card} J_h(y) \le (h+1)^n$, thus (3.6) yields $1 \le \psi_h^\sharp(y) \le \big(1+\frac{1}{h}\big)^n$ and we deduce that $\psi_h^\sharp$ converge pointwise to 1 on Y.

On the other hand, by (3.5) we have

$$\sup_{y\in Y} |\nabla\psi_h^\sharp(y)| \le \sup_{y\in Y} Ch^{-n}\big[\operatorname{card} J_h(y) - \operatorname{card} I_h(y)\big] \le Ch^{-n}\big[(h+1)^n - h^n\big] \ .$$

It follows that $\psi_h^\sharp$ and $\nabla\psi_h^\sharp$ both converge uniformly to 1 and 0, respectively, on Y, hence $\psi_h^\sharp$ converge to 1 strongly in $H^{1,\infty}_\mu(\mathbf{T})$. □

LEMMA 3.4. *For any $u \in \mathcal{C}^\infty(\mathbf{T})$ and any $\Phi \in X^{p'}_\mu(\mathbf{T})$, there holds*

$$\int_Y \nabla_\mu u \cdot \Phi \, d\mu = -\int_Y u \operatorname{div}_\mu \Phi \, d\mu \ . \tag{3.7}$$

Proof. Take a sequence $\{\psi_h\} \subset \mathcal{D}$ as in Lemma 3.3. For any $u \in \mathcal{C}^\infty(\mathbf{T})$ and for every h, the product function $u\psi_h$ belongs to $\mathcal{D}$. Then, for any $\Phi \in X^{p'}_\mu(\mathbf{T})$, we are allowed to write

$$\int_{\mathbb{R}^n} \nabla_\mu(u\psi_h) \cdot \Phi \, d\mu = -\int_{\mathbb{R}^n} u\psi_h \operatorname{div}_\mu \Phi \, d\mu. \tag{3.8}$$

Using Lemma 3.2, the left-hand side of the above equation can be rewritten as

$$\int_{\mathbb{R}^n} \big(\psi_h \nabla_\mu u + u\nabla_\mu\psi_h\big) \cdot \Phi \, d\mu = \int_Y \big(\psi_h^\sharp \nabla_\mu u + u\nabla_\mu\psi_h^\sharp\big) \cdot \Phi \, d\mu \ .$$

By the choice of $\{\psi_h\}$, (3.7) thus follows from (3.8) when passing to the limit as $h \to +\infty$. □

PROPOSITION 3.5. *The operator $A_\sharp$ is closable.*

Proof. We need to show that, if $\{u_h\} \subset \mathcal{C}^\infty(\mathbf{T})$ satisfies $u_h \to 0$ in $L^p_\mu(\mathbf{T})$ and $\nabla_\mu u_h \to v$ in $(L^p_\mu(\mathbf{T}))^n$, then $v = 0$ μ-a.e. on Y. To this aim, it is enough to show that

$$\sup\left\{\int_Y v\cdot\Phi\,d\mu \;:\; \Phi\in X^{p'}_\mu(\mathbf{T})\right\}=0. \tag{3.9}$$

Indeed, $X^{p'}_\mu(\mathbf{T})$ satisfies the following locality property: whenever $\Phi \in X^{p'}_\mu(\mathbf{T})$ and $\psi \in \mathcal{D}(Y)$, the field $\psi\Phi$ (extended by periodicity out of Y) belongs to $X^{p'}_\mu(\mathbf{T})$. Then a commutation argument between supremum and integral can be applied [8, Theorem 1], and, using also (3.3), (3.9) becomes

$$\int_Y \sup\{v\cdot z \;:\; z\in T_\mu(x)\}\,d\mu=0\ .$$

This, coupled with the information that $v(x) \in T_\mu(x)$ for μ-a.e. x (recall that $v = \lim_h \nabla_\mu u_h$), implies that v vanishes μ-a.e.

It remains to prove (3.9). To this aim, we use Lemma 3.4 and the convergence of $\{u_h\}$ to zero:

$$\int_Y v\cdot\Phi\,d\mu=\lim_{h\to+\infty}\int_Y\nabla_\mu u_h\cdot\Phi\,d\mu=-\int_Y u_h\,\mathrm{div}_\mu\,\Phi\,d\mu=0. \qquad \square$$

We can now give the following definition.

DEFINITION 3.6. *Set* $H^{1,p}_\mu(\mathbf{T}) := D(\overline{A}_\sharp)$, *and* $\nabla_\mu u := \overline{A}_\sharp u$ *for any* $u \in H^{1,p}_\mu(\mathbf{T})$.

It turns out by this definition that $H^{1,p}_\mu(\mathbf{T})$ is a closed subspace of $H^{1,p}_{\mu,\mathrm{loc}} := \{u \in L^p_{\mu,\mathrm{loc}} \;:\; u\psi \in H^{1,p}_\mu\ \forall\psi\in\mathcal{D}\}$ (in particular, it is reflexive for $p \in (1+\infty)$). Moreover, the integration by parts formula (3.7) can immediately be extended to any $u \in H^{1,p}_\mu(\mathbf{T})$.

Let us consider now the adjoint operator $A^*_\sharp$. We have

$$D(A^*_\sharp)=\left\{\sigma\in\left(L^{p'}_\mu(\mathbf{T})\right)^n \;:\; \left|\int_Y\sigma\cdot\nabla_\mu u\,d\mu\right|\le C\|u\|_{p',\mu}\ \forall u\in\mathcal{C}^\infty(\mathbf{T})\right\}=:Y^{p'}_\mu(\mathbf{T}).$$

PROPOSITION 3.7. $Y^{p'}_\mu(\mathbf{T}) = \left\{\sigma \in \left(L^{p'}_\mu(\mathbf{T})\right)^n \;:\; P_\mu\sigma \in X^{p'}_\mu(\mathbf{T})\right\}$; $A^*_\sharp\sigma = -\,\mathrm{div}_\mu(P_\mu\sigma)$.

Proof. Let us prove the first assertion. The inclusion $\supseteq$ is an immediate consequence of (3.7). For the converse, let $\sigma \in Y^{p'}_\mu(\mathbf{T})$, and let us show that the restriction $\mathrm{div}(P_\mu\sigma\mu)\llcorner\Omega$ of $\mathrm{div}(P_\mu\sigma\mu)$ to any bounded open set Ω belongs to $L^{p'}_\mu(\Omega)$. Take $\psi \in \mathcal{D}$ with $\mathrm{spt}\,\psi \subset \Omega$. Using Lemma 3.2, we get

$$-\langle\mathrm{div}(P_\mu\sigma\mu)\llcorner\Omega,\psi\rangle_{(\mathcal{D},\mathcal{D}')}=\int_{\mathbb{R}^n}P_\mu\sigma\cdot\nabla\psi\,d\mu=\int_Y P_\mu\sigma\cdot\nabla\psi^\sharp\,d\mu=\int_Y\sigma\cdot\nabla_\mu\psi^\sharp\,d\mu\ .$$

Then, since $\sigma \in Y^{p'}_\mu(\mathbf{T})$, there exists $C > 0$ such that

$$|\langle\mathrm{div}(P_\mu\sigma\mu)\llcorner\Omega,\psi\rangle|\le C\|\psi^\sharp\|_{p,\mu}\qquad\forall\psi\in\mathcal{D}(\Omega)\ . \tag{3.10}$$

Now, using the boundedness of Ω, we can find a positive constant C' such that

$$\|\psi^\sharp\|_{p,\mu,Y}\le C'\|\psi\|_{p,\mu,\Omega}\qquad\forall\psi\in\mathcal{D}(\Omega)\ . \tag{3.11}$$

Indeed, setting $I_\Omega := \{i \in \mathbf{Z}^n \,:\, (Y-i)\cap\Omega \neq \emptyset\}$, we obtain

$$\int_Y |\psi^\sharp|^p\,d\mu = \int_Y \left|\sum_{i\in I_\Omega}\psi(y-i)\right|^p d\mu \le \int_Y \left\{\sum_{i\in I_\Omega}|\psi(y-i)|\right\}^p d\mu$$
$$\le (\mathrm{card} I_\Omega)^{p-1}\sum_{i\in I_\Omega}\int_Y |\psi(y-i)|^p\,d\mu = (\mathrm{card} I_\Omega)^{p-1}\int_\Omega |\psi|^p\,d\mu.$$

Combining (3.10) and (3.11), we deduce that $\mathrm{div}(P_\mu\sigma\mu)\llcorner\Omega$ can be identified with an element of $L^{p'}_\mu(\Omega)$ and the first part of the statement is proved. To obtain the second assertion, we take $u\in\mathcal{C}^\infty(\mathbf{T})$ and $\sigma\in Y^{p'}_\mu(\mathbf{T})$. Then (3.7) yields

$$\int_Y \nabla_\mu u\cdot\sigma\,d\mu = \int_Y \nabla_\mu u\cdot P_\mu\sigma\,d\mu = -\int_Y u\,\mathrm{div}_\mu(P_\mu\sigma)\,d\mu\ . \qquad \square$$

As an immediate consequence of Proposition 3.7 and of Lemma 3.1, we can equivalently define $H^{1,p}_\mu(\mathbf{T})$ as follows:

$$H^{1,p}_\mu(\mathbf{T}) = \Big\{u\in L^p_\mu(\mathbf{T}) \text{ s.t. } \exists C>0 : |\langle u, \mathrm{div}(P_\mu\sigma\mu)\rangle| \le C\|\sigma\|_{p',\mu}\ \forall\sigma\in Y^{p'}_\mu(\mathbf{T})\Big\}. \tag{3.12}$$

The next theorem is the periodic version of Theorem 3.1 in [7], to which we refer for some details omitted in the proof.

PROPOSITION 3.8. *Let J be the functional defined on $L^p_\mu(\mathbf{T})$ $(p>1)$ by*

$$J(u) = \begin{cases} \displaystyle\int_Y j(y,\nabla u)\,d\mu & \text{if } u\in\mathcal{C}^\infty(\mathbf{T}),\\ +\infty & \text{if } u\in L^p_\mu(\mathbf{T})\setminus\mathcal{C}^\infty(\mathbf{T}),\end{cases}$$

where the integrand $j=j(y,z)$ is μ-measurable in y, convex in z, and satisfies, for some positive constants C,c, the growth condition

$$c|z|^p \le j(y,z) \le C\big(1+|z|^p\big) \qquad \forall (y,z)\in Y\times\mathbb{R}^n\ .$$

Then the relaxed functional of J on L^p_μ is given by

$$\overline{J}(u) = \begin{cases} \displaystyle\int_Y j_\mu(y,\nabla_\mu u)\,d\mu & \text{if } u\in H^{1,p}_\mu(\mathbf{T}),\\ +\infty & \text{if } u\in L^p_\mu(\mathbf{T})\setminus H^{1,p}_\mu(\mathbf{T}),\end{cases}$$

where

$$j_\mu(y,z) := \inf\big\{j(y,z+\xi)\ :\ \xi\in[T_\mu(y)]^\perp\big\}. \tag{3.13}$$

Remark 3.9. We stress that $j_\mu(y,z)$ depends only on the component of z along $T_\mu(y)$.

Proof. By convexity, $\overline{J}$ coincides with the bipolar functional J^{**} in the duality $\big(L^p_\mu(\mathbf{T}), L^{p'}_\mu(\mathbf{T})\big)$. Let $B: D(B)\subset L^p_\mu(\mathbf{T})\to\big(L^p_\mu(\mathbf{T})\big)^n$ be the linear operator defined by $D(B) := \mathcal{C}^\infty(\mathbf{T})$, $Bu=\nabla u$. Set $F_j(u) := \int_Y j(y,u)\,d\mu$. An abstract convex analysis lemma for the computation of $(F_j\circ B)^*$ gives (see [7, Theorem 5.1])

$$J^*(v) = (F_j\circ B)^*(v) = \inf\left\{\int_Y j^*(y,\Phi)\ :\ B^*\Phi = v\right\}\ , \qquad v\in L^{p'}_\mu(\mathbf{T})\ .$$

Now J^{**} equals by definition

$$J^{**}(u) = \sup\left\{\langle u, v\rangle - J^*(v) \ : \ v \in L^{p'}_\mu(\mathbf{T})\right\}$$

$$= \sup\left\{\langle u, B^*\Phi\rangle - \int_Y j^*(y,\Phi)\,d\mu \ : \ \Phi \in D(B^*)\right\}$$

It can be checked, arguing as in the proof of Proposition 3.7, that $D(B^*) = X^{p'}_\mu(\mathbf{T})$ and $B^*\Phi = -\operatorname{div}_\mu \Phi$, hence

$$J^{**}(u) = \sup\left\{\int_Y \left[-u\operatorname{div}_\mu\Phi - j^*(y,\Phi)\right] d\mu \ : \ \Phi \in X^{p'}_\mu(\mathbf{T})\right\}.$$

In particular, when $u \notin H^{1,p}_\mu(\mathbf{T})$, (3.12) and the growth condition satisfied by j^* lead to $J^{**}(u) = +\infty$. On the other hand, if $u \in H^{1,p}_\mu(\mathbf{T})$, we can integrate by parts and we obtain

$$J^{**}(u) = \sup\left\{\int_Y \left[\nabla_\mu u\cdot\Phi - j^*(y,\Phi)\right] d\mu \ : \ \Phi \in X^{p'}_\mu(\mathbf{T})\right\}.$$

Applying the same argument of commutation as used in the proof of Proposition 3.5, the supremum can be passed under the sign of integral; thus, taking into account (3.3), we obtain

$$\overline{J}(u) = J^{**}(u) = \int_Y \sup_{z^*\in T_\mu(y)} [\nabla_\mu u\cdot z^* - j^*(y,z^*)]d\mu = \int_Y j_\mu(y,\nabla_\mu u)d\mu\ . \qquad \square$$

4. Connectedness assumptions and two-scale limit of gradients. In this section, we consider a sequence $\{u_\varepsilon\} \subset \mathcal{C}^1_0(\Omega)$ which satisfies, as well as the sequence of its gradients $\{\nabla u_\varepsilon\}$, a uniform boundedness estimate in $L^p_{\mu_\varepsilon}$. In light of Proposition 2.3, we may associate to the pair $(u_\varepsilon, \nabla u_\varepsilon)$ a two-scale limit (u_0, χ). The aim of the structure result below (see Theorem 4.2) is to establish a differential relation between u_0 and χ, thus generalizing the Lebesgue case already studied in [1]. Dealing with a general periodic measure μ is considerably more delicate and requires us to introduce some notions of connectedness related to the Sobolev framework developed in section 3. Such notions, whose relationships are summarized in Remark 4.1 below, are all given for a fixed exponent $p \in [1, +\infty]$ and read as follows (C, c are supposed real constants).

- μ is *weakly p-connected on* $\mathbf{T}$ if

$$(H1) \qquad u \in H^{1,p}_\mu(\mathbf{T}), \quad \nabla_\mu u = 0 \ \mu\text{-a.e.} \quad \Rightarrow \quad \exists\, c \ : \ u = c \ \mu\text{-a.e.};$$

- μ is *weakly p-connected on* $\mathbb{R}^n$ if

$$(H2) \qquad u \in H^{1,p}_{\mu,loc}, \quad \nabla_\mu u = 0 \ \mu\text{-a.e.} \quad \Rightarrow \quad \exists\, c \ : \ u = c \ \mu\text{-a.e.};$$

- μ is *strongly p-connected on* $\mathbf{T}$ if

$$(H3) \qquad \exists\, C \ : \ \int_Y |u|^p\,d\mu \le C\int_Y |\nabla_\mu u|^p\,d\mu \quad \forall u \in H^{1,p}_\mu(\mathbf{T}) \text{ with } \int_Y u\,d\mu = 0;$$

- μ is *strongly p-connected on* $\mathbb{R}^n$ if

$$(H4) \qquad \exists\, C : \int_{kY} |u|^p\,d\mu \le Ck^p\int_{kY} |\nabla_\mu u|^p\,d\mu \ \ \forall k \in \mathbb{N}, \forall u \in H^{1,p}_\mu(k\mathbf{T}) \text{ with } \int_{kY} u\,d\mu = 0.$$

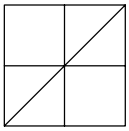


FIG. 4.1A.

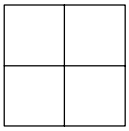


FIG. 4.1B.

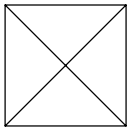


FIG. 4.1C.

Remark 4.1.

(i) We notice first that by density the properties $(H3)$–$(H4)$ need to be checked only on smooth functions. It is immediate that $(H4) \Rightarrow (H3) \Rightarrow (H1)$ and that $(H4) \Rightarrow (H2) \Rightarrow (H1)$. On the other hand we stress that $(H3) \not\Rightarrow (H2)$ (consider the case when $n = 3$ and μ is the measure $\mathcal{H}^1$ over a periodic network of e_3-parallel fibers) and that also $(H2) \not\Rightarrow (H3)$ (check by taking μ as the Lebesgue measure weighted by a suitable degenerated density).

(ii) The *assumption* $(H1)$. We stress that, in the special case when μ is the Lebesgue measure on an open periodic subset Q of $\mathbb{R}^n$, the connectedness of Q in the sense of [28, Definition 1.1] is equivalent to $(H1)$. Notice also that $(H1)$ does depend on p: in the well known-case of the Lebesgue measure concentrated on the black squares of a chess-board, it is satisfied if and only if $p > 2$. If, more drastically, the support of μ is not connected in the topological sense, then $(H1)$ is obviously false.

(iii) *The assumption* $(H4)$. Poincaré-type inequalities on manifolds or on weighted Sobolev spaces have recently received a deal of attention for their importance in different areas (see, for instance, [11], [22]). By $(H4)$ we require that the Poincaré constant on kY is finite for every $k \in \mathbb{N}$ and further that it is equal to k times the Poincaré constant on Y: by change of variables, this implies, in particular, that the Poincaré constant for each rescaled measure μ_ε on Y is independent of ε (see (5.4)). For instance, $(H4)$ is fulfilled if $n = 2$ and $\mu \llcorner Y$ is the one-dimensional Hausdorff measure over one of the sets described in Figure 4.1.

We can now prove the structure result mentioned in the opening of the section.

THEOREM 4.2. *Let $p > 1$, let Ω be an open bounded subset of $\mathbb{R}^n$ of class $\mathcal{C}^1$, and let $\{u_\varepsilon\} \subset \mathcal{C}_0^1(\Omega)$ satisfy $\int_\Omega \big(|u_\varepsilon|^p + |\nabla u_\varepsilon|^p\big)\, d\mu_\varepsilon \le M$; possibly passing to a subsequence, assume that $u_\varepsilon \rightharpoonup u_0 \in L^p_m(\Omega \times Y)$ and $\nabla u_\varepsilon \rightharpoonup \chi \in \big(L^p_m(\Omega \times Y)\big)^n$. Then*

(i) *if μ satisfies $(H1)$, $u_0(x, y) = u(x)$ (i.e. u_0 is independent of y), where the function u belongs to $W^{1,p}_0(\Omega)$ provided μ satisfies also $(H2)$ and $(H3)$;*

(ii) *under assumptions $(H2)$ and $(H3)$ on μ, there exists $u_1 \in L^p\big(\Omega, H^{1,p}_\mu(\mathbf{T})\big)$ such that $\chi(x, y) = \nabla u(x) + \nabla_{\mu,y} u_1(x, y) + \xi(y)$, with u as in* (i), *and $\xi(y) \in \big[T_\mu(y)\big]^\perp$ μ-a.e. on Y; in addition, $\nabla_{\mu_\varepsilon} u_\varepsilon \rightharpoonup \nabla u(x) + \nabla_{\mu,y} u_1(x, y)$.*

The proof of Theorem 4.2 requires some intermediate steps, which are given in the form of autonomous lemmas.

LEMMA 4.3. *Let $V := \big\{ \operatorname{div}_\mu \Phi \,:\, \Phi \in X^{p'}_\mu(\mathbf{T}) \big\}$. If μ enjoys $(H1)$, then the orthogonal space $V^\perp$ of V in $L^p_\mu(\mathbf{T})$ is given by the constant functions.*

Remark 4.4. A consequence of the above lemma is that the closure of V in $L^{p'}_\mu(\mathbf{T})$ is given by the functions with zero mean value. A similar result has been obtained in a different framework by Zhikov in the case $p = 2$ (see [28]).

Proof of Lemma 4.3. Consider the functional J on $L^p_\mu(\mathbf{T})$:

$$J(u) = \begin{cases} \dfrac{1}{p} \displaystyle\int_Y |\nabla_\mu u|^p \, d\mu & \text{if } u \in H^{1,p}_\mu(\mathbf{T}), \\ +\infty & \text{if } u \in L^p_\mu(\mathbf{T}) \setminus H^{1,p}_\mu(\mathbf{T}). \end{cases}$$

We can write J as $J = F \circ A_\sharp$, where $F(z) := \frac{1}{p}\int |z|^p \, d\mu$, and the linear operator $A_\sharp : D(A_\sharp) \subset L^p_\mu(\mathbf{T}) \to \big(L^p_\mu(\mathbf{T})\big)^n$ is defined as in section 3. For the computation of $(F \circ A_\sharp)^*$, we can apply the same lemma already used in the proof of Proposition 3.8 (see [7, Theorem 5.1]), which gives, for any $v \in L^{p'}_\mu(\mathbf{T})$,

$$\begin{aligned} J^*(v) &= \inf\left\{ \frac{1}{p'}\int_Y |\sigma|^{p'} \, d\mu \ : \ A^*_\sharp \sigma = v \right\} \\ &= \inf\left\{ \frac{1}{p'}\int_Y |\sigma|^{p'} \, d\mu \ : \ \sigma \in Y^{p'}_\mu(\mathbf{T}) \ , \ -\mathrm{div}_\mu(P_\mu \sigma) = v \right\} \\ &= \inf\left\{ \frac{1}{p'}\int_Y |\sigma|^{p'} \, d\mu \ : \ \sigma \in X^{p'}_\mu(\mathbf{T}) \ , \ -\mathrm{div}_\mu(\sigma) = v \right\}. \end{aligned}$$

Hence V coincides with $\mathrm{dom}(J^*) := \big\{ v \in L^{p'}_\mu(\mathbf{T}) \ : \ J^*(v) < +\infty \big\}$. Let J^∞ be the recession function of J. It is defined by

$$J^\infty(u) = \lim_{t\to+\infty} \frac{J(tu)}{t} = \begin{cases} +\infty & \text{if } u \notin H^{1,p}_\mu(\mathbf{T}), \\ \\ \displaystyle\lim_{t\to+\infty} \frac{t^{p-1}}{p}\int_Y |\nabla_\mu u|^p \, d\mu & \text{if } u \in H^{1,p}_\mu(\mathbf{T}). \end{cases}$$

Due to the assumption $(H1)$ on μ, we get

$$J^\infty(u) < +\infty \iff J^\infty(u) = 0 \iff \exists\, c \ : \ u = c \ \mu - \text{a.e. on } Y \ . \tag{4.1}$$

As J is convex, lower semicontinuous, and proper ($J(0) = 0$), by a well-known result in convex analysis (see, for instance, [27, Theorem 13.3]), the recession functional J^∞ of J satisfies

$$J^\infty(u) = \sup\left\{ \int_Y u\, v \, d\mu \ : \ v \in \mathrm{dom}(J^*) \right\} \ , \qquad u \in L^p_\mu(\mathbf{T}).$$

Thus u belongs to the othogonal of $V = \mathrm{dom}(J^*)$ if and only if $J^\infty(u) = 0$. The conclusion follows from (4.1). □

LEMMA 4.5. *Let j satisfy the assumptions of Proposition* 3.8*, and set*

$$j^{\mathrm{hom}}(z) := \inf\left\{ \int_Y j(y, z + \nabla u(y)) \, d\mu \ : \ u \in \mathcal{C}^\infty(\mathbf{T}) \right\}, \qquad z \in \mathbb{R}^n. \tag{4.2}$$

Then

(i) *j^{hom} is convex and satisfies a growth condition $j^{\mathrm{hom}}(z) \le \Lambda\big(1 + |z|^p\big)$ ($\Lambda \in (0, +\infty)$);*

(ii) *$\big(j^{\mathrm{hom}}\big)^*(z^*) = \inf\Big\{ \int_Y j^*(y, \Phi) \, d\mu \ : \ \Phi \in X^{p'}_\mu(\mathbf{T}),\ \mathrm{div}_\mu \Phi = 0,\ \int_Y \Phi \, d\mu = z^* \Big\}$;*

(iii) *$j^{\mathrm{hom}}(z) = \inf\Big\{ \int_Y j_\mu(y, z_\mu + \nabla_\mu u(y)) \, d\mu \ : \ u \in H^{1,p}_\mu(\mathbf{T}) \Big\}$, where $z_\mu := P_\mu(y)[z]$;*

(iv) *under the hypotheses $(H2)$ and $(H3)$ on μ, j^{hom} enjoys a coercivity property $j^{\mathrm{hom}}(z) \ge \lambda |z|^p$ ($\lambda \in (0, +\infty)$), hence in particular $(j^{\mathrm{hom}})^*$ is continuous on the whole $\mathbb{R}^n$.*

Proof. The convexity and the growth assumption on j give straightforwardly the corresponding properties for j^{hom}.

To compute $(j^{\rm hom})^*$ (in the duality $\big(\big(L^p_\mu(\mathbf{T})\big)^n, \big(L^{p'}_\mu(\mathbf{T})\big)^n\big)$), notice that $j^{\rm hom}$ can be written as

$$j^{\rm hom} = F_j \nabla\chi_G + \chi_C \ , \tag{4.3}$$

where $F_j(u) = \int_Y j(y,u)\,d\mu$, $\chi_G(u)$ and $\chi_C(u)$ are finite and equal to zero if and only if u belongs, respectively, to

$$G := \big\{\nabla\psi \ : \ \psi \in \mathcal{C}^\infty(\mathbf{T})\big\}\,, \quad C := \big\{v \in (L^p_\mu(\mathbf{T}))^n \text{ s.t. } \exists\, z \in \mathbb{R}^n \ : \ v = z \ \mu\text{-a.e. on } Y\big\}\,,$$

and $(F_j\nabla\chi_G)(u) := \inf\big\{F_j(u+v) + \chi_G(-v) \ : \ v \in \big(L^p_\mu(\mathbf{T})\big)^n\big\}$. Being F_j continuous at $0 \in \mathrm{dom}(\chi_G)$, we deduce from (4.3) (see, for instance, [27, Theorem 16.4])

$$(j^{\rm hom^*}) = (F_j\nabla\chi_G)^*\nabla\chi^*_C = \big(F^*_j + \chi^*_G\big)\nabla\chi^*_C \ .$$

We have $\chi^*_G = \chi_{G^\perp}$, $\chi^*_C = \chi_{C^\perp}$, with (recall Lemma 3.4)

$$G^\perp = \big\{\Phi \in X^{p'}_\mu(\mathbf{T}) \ : \ \mathrm{div}_\mu\,\Phi = 0\big\}\,, \qquad C^\perp = \left\{v \in \big(L^{p'}_\mu(\mathbf{T})\big)^n \ : \ \int_Y v\,d\mu = 0\right\}.$$

Thus we get, using [13, Theorem VII.14] for the computation of F^*_j,

$$\begin{aligned}
\big(j^{\rm hom}\big)^*(z^*) &= \inf\big\{F^*_j(z^*+\sigma) \ : \ z^*+\sigma \in G^\perp \ , \ \sigma \in C^\perp\big\} \\
&= \inf\left\{\int_Y j^*(y, z^*+\sigma)\,d\mu : z^*+\sigma \in X^{p'}_\mu(\mathbf{T}),\, \mathrm{div}_\mu(z^*+\sigma) = 0,\right. \\
&\qquad\qquad \left.\int_Y \sigma\,d\mu = 0\right\} \\
&= \inf\left\{\int_Y j^*(y,\Phi)\,d\mu \ : \ \Phi \in X^{p'}_\mu(\mathbf{T}),\ \mathrm{div}_\mu\,\Phi = 0,\ \int_Y \Phi\,d\mu = z^*\right\}.
\end{aligned}$$

The assertion (iii) is a consequence of Proposition 3.8 and of a well-known property of relaxation (see [12, Chapter 1]).

We finally pass to prove (iv). The growth condition from below on j implies

$$j^{\rm hom}(z) \geq c \inf\left\{\int_Y |z_\mu + \nabla_\mu u|^p\,d\mu \ : \ u \in H^{1,p}_\mu(\mathbf{T})\right\} =: c f(z).$$

In fact, by the assumption (H3) on μ, the infimum above is attained on $H^{1,p}_\mu(\mathbf{T})$. Then, we claim that $f(z) > 0$ whatever $z \neq 0$. Indeed, if $f(z) = 0$, there exists $u \in H^{1,p}_\mu(\mathbf{T})$ such that $z_\mu + \nabla_\mu u = 0$ μ-a.e. on Y, or equivalently $\nabla_\mu(z\cdot y + u) = 0$ μ-a.e. on Y. Since the function $y \mapsto z\cdot y + u\,(x)$ belongs to $H^{1,p}_{\mu,\rm loc}$, the assumption $(H2)$ on μ implies that it must be constant μ-a.e., and this is equivalent, by the periodicity of u, to requiring that z equals zero. Denoting by m_1 the (strictly positive) minimum of f on the unit ball of $\mathbb{R}^n$, and using the p-homogeneity of f, we deduce that the coercivity property (iv) holds with $\lambda = cm_1$. □

Lemma 4.6. *Let $M := \big\{\sigma \in Y^{p'}_\mu(\mathbf{T}) \ : \ \mathrm{div}_\mu(P_\mu\sigma) = 0\big\}$. If μ enjoys $(H3)$, then the orthogonal space $M^\perp$ of M in $\big(L^{p'}_\mu(\mathbf{T})\big)^n$ is given by $N := \big\{\nabla_\mu u \ : \ u \in H^{1,p}_\mu(\mathbf{T})\big\}$.*

Proof. We have $N = R(A_\sharp) :=$ image of $A_\sharp$, and $M = \ker(A^*_\sharp)$, $A^*_\sharp$ being the linear operator defined in section 3. Therefore $M = N^\perp$ [10, Corollary II.17], and so

$M^\perp = \overline{N}$. It remains to prove that N is closed in $\big(L^p_\mu(\mathbf{T})\big)^n$. To this aim, we exploit the assumption $(H3)$ on μ. Let $\{\nabla_\mu u_h\}$ be a sequence of elements of N converging in $\big(L^p_\mu(\mathbf{T})\big)^n$. Then $\big\{u_h - \int_Y u_h\,d\mu\big\}$ is a Cauchy sequence in $H^{1,p}_\mu(\mathbf{T})$, since by $(H3)$

$$\left\|u_h - \int_Y u_h\,d\mu - u_k + \int_Y u_k\,d\mu\right\|_{p,\mu} \le \big\|\nabla_\mu u_h - \nabla_\mu u_k\big\|_{p,\mu}.$$

By completeness of $H^{1,p}_\mu(\mathbf{T})$, the sequence $\big\{u_h - \int_Y u_h\,d\mu\big\}$ converges to some $u \in H^{1,p}_\mu(\mathbf{T})$, so that the limit of $\{\nabla_\mu u_h\}$ equals $\nabla_\mu u$ and in particular it belongs to N. □

We can now give the following.

Proof of Theorem 4.2. Let $\Phi \in X^{p'}_\mu(\mathbf{T})$ and $\psi \in \mathcal{D}(\overline{\Omega})$. Using (2.1), one can check that $\operatorname{div}\big(\Phi(\frac{x}{\varepsilon})\mu_\varepsilon\big) = \varepsilon^{-1}(\operatorname{div}_\mu \Phi)\big(\frac{x}{\varepsilon}\big)\mu_\varepsilon$. Then (3.1) gives

$$\operatorname{div}\left(\psi(x)\Phi\left(\frac{x}{\varepsilon}\right)\mu_\varepsilon\right) = \nabla\psi(x)\cdot\Phi\left(\frac{x}{\varepsilon}\right)\mu_\varepsilon + \varepsilon^{-1}\psi(x)(\operatorname{div}_\mu\Phi)\left(\frac{x}{\varepsilon}\right)\mu_\varepsilon. \tag{4.4}$$

Since u_ε belong to $\mathcal{D}$ (when extended to zero out of Ω), both sides of the above inequality can be applied to u_ε, and, multiplying by ε, we get

$$\begin{aligned}\varepsilon\int_\Omega \psi(x)\nabla u_\varepsilon(x)\cdot\Phi\left(\frac{x}{\varepsilon}\right)d\mu_\varepsilon \\ = \ \varepsilon\int_\Omega u_\varepsilon(x)\nabla\psi(x)\cdot\Phi\left(\frac{x}{\varepsilon}\right)d\mu_\varepsilon + \int_\Omega u_\varepsilon(x)\psi(x)(\operatorname{div}_\mu\Phi)\left(\frac{x}{\varepsilon}\right)d\mu_\varepsilon.\end{aligned}$$

If we pass to the limit as $\varepsilon\to 0$, taking into account that $\{u_\varepsilon\}$ and $\{\nabla u_\varepsilon\}$ are both uniformly bounded in $L^p_{\mu_\varepsilon}(\Omega)$, we obtain

$$\lim_{\varepsilon\to 0}\int_\Omega u_\varepsilon(x)\psi(x)(\operatorname{div}_\mu\Phi)\left(\frac{x}{\varepsilon}\right)d\mu_\varepsilon = 0.$$

Now, in view of Remark 2.4, $\varphi(x,y) := \psi(x)\operatorname{div}_\mu\Phi(y)$ is an admissible test function with respect to the two-scale convergence, and we infer

$$\int_\Omega \psi(x)\left[\int_Y u_0(x,y)\operatorname{div}_\mu\Phi(y)\,d\mu(y)\right]dx = 0.$$

By the arbitrariness of $\psi\in\mathcal{D}(\overline{\Omega})$ and of $\Phi\in X^{p'}_\mu(\mathbf{T})$, it follows that, for $\mathcal{L}^n$-a.e. $x\in\Omega$, $u_0(x,\cdot)$ belongs to $V^\perp$ with V defined as in Lemma 4.3. Then by the same lemma we deduce that, for $\mathcal{L}^n$-a.e. $x\in\Omega$, $u_0(x,\cdot)$ is constant μ-a.e. on Y.

Set $u(x) := u_0(x,y)$, and let us prove that u belongs to $W^{1,p}_0(\Omega)$ under the assumptions $(H2)$ and $(H3)$ on μ. Let ψ, Φ be as above, and assume in addition that $\operatorname{div}_\mu\Phi = 0$. Applying both sides of (4.4) to u_ε (this time without multiplying by ε), it follows that

$$-\int_\Omega \psi(x)\nabla u_\varepsilon(x)\cdot\Phi\left(\frac{x}{\varepsilon}\right)d\mu_\varepsilon = \int_\Omega u_\varepsilon(x)\nabla\psi(x)\cdot\Phi\left(\frac{x}{\varepsilon}\right)d\mu_\varepsilon.$$

Passing to the two-scale limit, we deduce

$$\begin{aligned}-\int_{\Omega\times Y}\psi(x)\chi(x,y)\cdot\Phi(y)\,dm &= \int_{\Omega\times Y} u(x)\nabla\psi(x)\cdot\Phi(y)\,dm \\ &= \left\{\int_\Omega u(x)\nabla\psi(x)\,dx\right\}\cdot\left\{\int_Y\Phi(y)\,d\mu\right\}.\end{aligned} \tag{4.5}$$

By the Hölder inequality, it follows that

$$\overline{\Phi}\cdot\int_\Omega u\nabla\psi\,dx\le\|\chi\|_{p,m,\Omega\times Y}\|\Phi\|_{p',\mu,Y}\|\psi\|_{p',\mathcal{L}^n,\Omega}\ , \tag{4.6}$$

where we have set $\overline{\Phi}:=\int_Y\Phi(y)\,d\mu(y)$. Consider now the convex subset of $\mathbb{R}^n$

$$K:=\Big\{\overline{\Phi}\ :\ \Phi\in X_\mu^{p'}\ ,\ \mathrm{div}_\mu\,\Phi=0\ ,\ \|\Phi\|_{p',\mu,Y}\le 1\Big\}\ . \tag{4.7}$$

Notice that, by Lemma 4.5 (iii) and the definition of K, we have

$$\mathrm{dom}\,\big(j^{\mathrm{hom}}\big)^*=\bigcup_{\lambda\in\mathbb{R}}\lambda K\ .$$

If μ satisfies $(H2)$ and $(H3)$, we know from Lemma 4.5(iv) that $\mathrm{dom}\,\big(j^{\mathrm{hom}}\big)^*=\mathbb{R}^n$, so K must have a nonempty interior. By the regularity assumption on Ω, it follows from (4.6) that u belongs to $W_0^{1,p}(\Omega)$ (cf. [10, Proposition IX.18]).

We can in this case integrate by parts on Ω at the right-hand side of (4.5). We get

$$\int_\Omega\psi(x)\left\{\int_Y\big[\chi(x,y)-\nabla u(x)\big]\cdot\Phi(y)\,d\mu(y)\right\}dx=0.$$

By the arbitrariness of $\psi\in\mathcal{D}(\overline{\Omega})$, it follows that, for $\mathcal{L}^n$-a.e.$x\in\Omega$ and for every vector field $\Phi\in X_\mu^{p'}(\mathbf{T})$ with $\mathrm{div}_\mu\,\Phi=0$, there holds

$$\int_Y\big[\chi(x,y)-\nabla u(x)\big]\cdot\Phi(y)\,d\mu(y)=0\ .$$

In particular, taking $\Phi=P_\mu\sigma$, with $\sigma\in Y_\mu^{p'}(\mathbf{T})$, we deduce that $P_\mu(\cdot)\big[\chi(x,\cdot)-\nabla u(x)\big]$ belongs to $M^\perp$, where M is defined as in Lemma 4.6. Using such lemma, it then follows that there exists $u_1\in L^p(\Omega,H_\mu^{1,p}(\mathbf{T}))$ such that $P_\mu(y)\big[\chi(x,y)-\nabla u(x)\big]=\nabla_{\mu,y}u_1(x,y)$. Hence $\chi(x,y)=\nabla u(x)+\nabla_{\mu,y}u_1(x,y)+\xi(y)$, with $\xi(y)\in[T_\mu(y)]^\perp$. Finally, we apply Proposition 2.8 to the sequences $P_{\mu_\varepsilon}\in L^{p'}_{\mu_\varepsilon}(\Omega;\mathbb{R}^{n^2})$ and $\nabla u_\varepsilon\in L^p_{\mu_\varepsilon}(\Omega;\mathbb{R}^n)$: we have, respectively, $P_{\mu_\varepsilon}\twoheadrightarrow P_\mu(y)$ (see Example 2.7) and $\nabla u_\varepsilon\rightharpoonup\chi(x,y)$, so that the product $P_{\mu_\varepsilon}(\nabla u_\varepsilon)=\nabla_{\mu_\varepsilon}u_\varepsilon$ two-scale converges to $P_\mu(y)\big(\chi(x,y)\big)=\nabla u(x)+\nabla_{\mu,y}u_1(x,y)$. □

Remark 4.7. When $(H2)$ does not hold, we can apply all the steps of the above proof except that in this case the convex set K defined by (4.7) has an empty interior. Thus, denoting by M the subspace of $\mathbb{R}^n$ spanned by the relative interior of K, the estimate $z\cdot\nabla u\in L^p(\Omega)$ can be obtained only for $z\in M$. Therefore, in part (i) of Theorem 4.2, the assertion $u\in W_0^{1,p}(\Omega)$ must be replaced by $u\in W^{1,p}_{0,M}(\Omega)$, where

$$W^{1,p}_{0,M}(\Omega):=\big\{u\in L^p(\Omega)\ :\ \forall z\in M\ ,\ \ z\cdot\nabla u\in L^p(\Omega)\ \text{ and }\ u\,(\nu_\Omega\cdot z)=0\ \text{on}\ \partial\Omega\big\}.$$

5. Homogenization with periodic measures. We can now consider the homogenization problem introduced at the beginning of the paper, namely, the study of the Γ-convergence in the sense of De Giorgi [19] of the sequence of functionals

$$J_\varepsilon(u)=\begin{cases}\displaystyle\int_\Omega j\big(\frac{x}{\varepsilon},\nabla u\big)\,d\mu_\varepsilon & \text{if } u\in\mathcal{C}_0^1(\Omega),\\ +\infty & \text{if } u\in L^p_\mu(\Omega)\setminus\mathcal{C}_0^1(\Omega).\end{cases} \tag{5.1}$$

Let us fix here the assumptions made on the integrand $j(y,z)$, which, however, fall into the standard setting already employed in section 3: j is μ-measurable and Y-periodic in y, convex in z, and satisfies, for suitable positive constants C, c, the p-growth condition (where $p > 1$)

$$c|z|^p \le j(y,z) \le C\big(1+|z|^p\big)\ , \qquad (y,z) \in \mathbb{R}^n \times \mathbb{R}^n. \tag{5.2}$$

Naturally, the first step to be done in the study of the asymptotic behavior of J_ε is the choice of a suitable convergence on the class of all admissible functions u. To this purpose, note that the role played by u in the expression of $J_\varepsilon(u)$ is that of a function defined μ_ε-a.e.; thus, as μ_ε varies with ε, it is convenient to see the convergence of u_ε to u as the weak star convergence of the measures μ_ε to $u\mathcal{L}^n$. Therefore we implicitly extend the definition (5.1) of J_ε to the space $\mathcal{M}$ of Radon measures on $\mathbb{R}^n$ by setting

$$J_\varepsilon(\lambda) = \begin{cases} \displaystyle\int_\Omega j\big(\frac{x}{\varepsilon}, \nabla u\big)\, d\mu_\varepsilon & \text{if } \lambda = u\mu_\varepsilon\ ,\ u \in \mathcal{C}_0^1(\Omega), \\ +\infty & \text{otherwise.} \end{cases} \tag{5.3}$$

The equicoerciveness of (J_ε) on $\mathcal{M}$ obtained in Lemma 5.1 below is related to the strong p-connectness property of μ.

LEMMA 5.1. *Let $p > 1$ and μ be a periodic Radon measure satisfying $(H4)$. Then, for any sequence $\{u_\varepsilon\} \in \mathcal{C}_0^1(\Omega)$ such that $J_\varepsilon(u_\varepsilon\mu_\varepsilon)$ is upper bounded, we have that $\sup_\varepsilon \int_\Omega |u_\varepsilon|^p\, d\mu_\varepsilon < +\infty$; therefore $\{u_\varepsilon\mu_\varepsilon\}$ is relatively compact in $\mathcal{M}$ and any of its limit points is absolutely continuous with respect to $\mathcal{L}^n$, with a density $u \in L^p(\Omega)$.*

Proof. By the coerciveness assumption on j, there exists $M' > 0$ such that $\int_\Omega |\nabla u_\varepsilon|^p\, d\mu_\varepsilon \le M'$. Notice now that, for any $\varepsilon > 0$, any integer k, and any $v \in \mathcal{C}_0^1(\varepsilon k Y)$, there holds

$$\int_{\varepsilon k Y} |v|^p\, d\mu_\varepsilon \le (\varepsilon k C)^p \int_{\varepsilon k Y} |\nabla v|^p\, d\mu_\varepsilon, \tag{5.4}$$

where C is the positive and finite constant involved in $(H4)$. Indeed, using the definition of μ_ε and the assumption $(H4)$ on μ, we get

$$\begin{aligned} \int_{\varepsilon k Y} |v|^p\, d\mu_\varepsilon &= \varepsilon^n \int_{kY} |v(\varepsilon x)|^p\, d\mu \le \varepsilon^n (kC)^p \int_{kY} \big|\nabla\big(v(\varepsilon x)\big)\big|^p\, d\mu \\ &= \varepsilon^n (\varepsilon k C)^p \int_{kY} |\nabla v(\varepsilon x)|^p\, d\mu = (\varepsilon k C)^p \int_{\varepsilon k Y} |\nabla v|^p\, d\mu_\varepsilon\ . \end{aligned}$$

Since Ω is bounded, for any $\varepsilon > 0$ there exists a positive integer k_ε, with $\varepsilon k_\varepsilon \le C' < +\infty$, such that $\Omega \subset \varepsilon k_\varepsilon Y$. Applying (5.4) to every $u_\varepsilon \in \mathcal{C}_0^1(\Omega) \subset \mathcal{C}_0^1(\varepsilon k_\varepsilon Y)$, we infer

$$\int_\Omega |u_\varepsilon|^p\, d\mu_\varepsilon = \int_{\varepsilon k_\varepsilon Y} |u_\varepsilon|^p\, d\mu_\varepsilon \le (C')^p C^p \int_{\varepsilon k_\varepsilon Y} |\nabla u_\varepsilon|^p\, d\mu_\varepsilon \le (C')^p C^p M'\ . \tag{5.5}$$

Then the weak compactness of $\{u_\varepsilon\mu_\varepsilon\}$ in $\mathcal{M}$, and the fact that any of its limit points is absolutely continuous with respect to $\mathcal{L}^n$, with a density $u \in L^p(\mathbb{R}^n)$, follows straightforward from Proposition 2.3. ∎

We can now state our main Γ-convergence theorem. It extends to general periodic measures μ the well-known result proved for the first time by Marcellini [24] in the case when μ is the Lebesgue measure over the complement of a periodic system of balls.

For the sake of completeness, we preliminarily recall that the the sequence $\{J_\varepsilon\}$ Γ-converges to $J^{\rm hom}$ if and only if both of the following inequalities hold:

(I) $\inf\Big\{ \liminf_{\varepsilon\to 0} J_\varepsilon(\lambda_\varepsilon) \ : \ \lambda_\varepsilon \rightharpoonup \lambda\Big\} \geq J^{\rm hom}(\lambda)\ , \quad \lambda\in\mathcal{M}\ ;$

(II) $\inf\Big\{ \limsup_{\varepsilon\to 0} J_\varepsilon(\lambda_\varepsilon) \ : \ \lambda_\varepsilon \rightharpoonup \lambda\Big\} \leq J^{\rm hom}(\lambda)\ , \quad \lambda\in\mathcal{M}\ .$

THEOREM 5.2. *Let μ satisfy $(H4)$ (and let $p>1$). Then the sequence $\{J_\varepsilon\}$ defined in* (5.3) *Γ-converges on $\mathcal{M}$ as $\varepsilon\to 0$ to the homogenized functional $J^{\rm hom}$ defined by*

$$J^{\rm hom}(\lambda) = \begin{cases} \displaystyle\int_\Omega j^{\rm hom}(\nabla u(x))\,dx & \text{if } \lambda = u\mathcal{L}^n\ ,\ u\in W^{1,p}_0(\Omega),\\ +\infty & \text{otherwise,}\end{cases} \tag{5.6}$$

where the integrand $j^{\rm hom}$ is defined via the unit-cell problem (4.2).

Remark 5.3. Theorem 5.2 can also be formulated replacing $(H4)$ by $(H2)$–$(H3)$: in this case we need to restrict ourselves to sequences $\{u_\varepsilon\}$ such that $\sup_\varepsilon \int |u_\varepsilon|^p\,d\mu_\varepsilon < +\infty$. One can also skip the assumption $(H2)$: in this case the integrand $j^{\rm hom}$ is still given by formula (4.2), but it degenerates in the directions $z\in M^\perp$, where M is the linear space defined in Remark 4.7; accordingly, $W^{1,p}_0(\Omega)$ has to be replaced by $W^{1,p}_{0,M}$ in the homogenization formula (5.6).

Proof of Theorem 5.2. We divide the proof into two parts, showing separately that the inequalities (I) and (II) required for the Γ-convergence hold.

Proof of (I). By Lemma 5.1, it is enough to prove (I) for $\lambda = u\mathcal{L}^n$, with $u\in W^{1,p}_0(\Omega)$. Let $u_\varepsilon\mu_\varepsilon \rightharpoonup u\mathcal{L}^n$, with $J_\varepsilon(u_\varepsilon\mu_\varepsilon)\leq M$. Then, by Lemma 5.1 and Theorem 4.2(i), we can assume that $\{u_\varepsilon\}$ two-scale converges to a function $u_0(x,y)$ independent of y, which then must be equal to $u(x)$. Moreover, we know from Theorem 4.2(ii) that $\nabla u_\varepsilon \rightharpoonup \chi$, where $\chi(x,y) = \nabla u(x) + \nabla_{\mu,y} u_1(x,y) + \xi(y)$, with $u_1\in L^p(\Omega, H^{1,p}_\mu(\mathbf{T}))$ and $\xi(y)\in \big[T_\mu(y)\big]^\perp$ μ-a.e. Applying Proposition 2.5, definition (3.13) of j_μ, and (iii) of Lemma 4.5, we obtain

$$\begin{aligned}\liminf_{\varepsilon\to 0} J_\varepsilon(u_\varepsilon\mu_\varepsilon) &\geq \int_{\Omega\times Y} j(y,\chi(x,y))\,dm = \int_{\Omega\times Y} j\big(y,\nabla u(x)+\nabla_{\mu,y}u_1(x,y)+\xi(y)\big)\,dm\\ &\geq \int_\Omega \left\{\int_Y \big[j_\mu(y,\nabla u(x)+\nabla_{\mu,y}u_1(x,y)\big]\,d\mu\right\} dx \geq \int_\Omega j^{\rm hom}\big(\nabla u(x)\big)\,dx.\end{aligned}$$

Proof of (II). Let $\lambda = u\mathcal{L}^n$, with $u\in W^{1,p}_0(\Omega)$; otherwise, there is nothing to prove. We have to find $\{u_\varepsilon\}\subset \mathcal{C}^1_0(\Omega)$ such that $\limsup_{\varepsilon\to 0} J_\varepsilon(u_\varepsilon)\leq J^{\rm hom}(\lambda)$. In view of the density of $\mathcal{D}(\Omega)$ into $W^{1,p}(\Omega)$, and of the continuity of $J^{\rm hom}$, by a standard diagonalization argument it is not restrictive to assume that $u\in\mathcal{D}(\Omega)$. For an arbitrary function $\varphi\in\mathcal{D}(\Omega;\mathcal{C}^\infty(\mathbf{T}))$, consider the sequence $\{u_\varepsilon\}\subset\mathcal{D}(\Omega)$ defined by

$$u_\varepsilon(x) = u(x) + \varepsilon\varphi\left(x,\frac{x}{\varepsilon}\right).$$

The measures $\{u_\varepsilon\mu_\varepsilon\}$ converge weakly to $u\mathcal{L}^n$, because $\varphi\big(x,\frac{x}{\varepsilon}\big)\mu_\varepsilon \rightharpoonup \{\int_Y \varphi(x,y)\,d\mu(y)\}\mathcal{L}^n$, and so $\varepsilon\varphi\big(x,\frac{x}{\varepsilon}\big)\mu_\varepsilon \rightharpoonup 0$. We have

$$\nabla u_\varepsilon(x) = \nabla u(x) + \varepsilon\nabla_x\varphi\left(x,\frac{x}{\varepsilon}\right) + \nabla_y\varphi\left(x,\frac{x}{\varepsilon}\right);$$

hence, by a well-known lemma concerning convex functions which satisfy a p-growth condition (see, for instance, [20, eq. 4.2.1]), the following inequality holds for some

positive constant C:

$$\begin{aligned}&\left|j(\nabla u_\varepsilon(x)) - j\left(\nabla u(x) + \nabla_y\varphi\left(x,\frac{x}{\varepsilon}\right)\right)\right| \\ &\qquad \le C\varepsilon\left|\nabla_x\varphi\left(x,\frac{x}{\varepsilon}\right)\right|\left\{1 + \left|\nabla u(x)\right|^{p-1} + \left|\nabla_y\varphi\big(x,\frac{x}{\varepsilon}\big)\right|^{p-1}\right\}.\end{aligned}$$

Integrating over Ω, a quite standard application of the Hölder inequality yields

$$\lim_{\varepsilon\to 0}\int_\Omega j(\nabla u_\varepsilon)\,d\mu_\varepsilon = \lim_{\varepsilon\to 0}\int_\Omega j\left(\nabla u(x) + \nabla_y\varphi\left(x,\frac{x}{\varepsilon}\right)\right)\,d\mu_\varepsilon.$$

Hence, to compute the limit of $J_\varepsilon(u_\varepsilon)$, we are reduced to having to apply convergence (2.3) to the right-hand side of the above equality choosing $v(x,y) = j\big(\nabla u(x) + \nabla_y\varphi(x,y)\big)$ as a test function. So we infer

$$\lim_{\varepsilon\to 0}\int_\Omega j\left(\frac{x}{\varepsilon},\nabla u_\varepsilon\right)\,d\mu_\varepsilon = \int_{\Omega\times Y} j\big(y,\nabla u(x) + \nabla_y\varphi(x,y)\big)\,dm.$$

In particular, we deduce that the upper Γ-limit of J_ε, defined by

$$J(u) := \inf_{u_\varepsilon\mu_\varepsilon \rightharpoonup u\mathcal{L}^n}\left\{\limsup_{\varepsilon\to 0}\int_\Omega j\left(\frac{x}{\varepsilon},\nabla u_\varepsilon\right)\,d\mu_\varepsilon\right\},$$

satisfies

$$J(u) \le \inf_{\varphi\,\in\,\mathcal{D}(\Omega;\mathcal{C}^\infty(\mathbf{T}))}\left\{\int_{\Omega\times Y} h_\varphi(x)\,dm\right\},$$

where

$$h_\varphi(x) := \int_Y j\big(y,\nabla u(x) + \nabla_y\varphi(x,y)\big)\,d\mu(y).$$

Noticing that, by a standard localization argument, we have for all $x\in\Omega$

$$\inf_{\varphi\,\in\,\mathcal{D}(\Omega;\mathcal{C}^\infty(\mathbf{T}))} h_\varphi(x) = \inf_{\varphi\,\in\,\mathcal{C}^\infty(\mathbf{T})} h_\varphi(x) = j^{hom}(\nabla u(x)),$$

the proof of the inequality $J(u)\le J^{hom}(u)$ (i.e., of (II)) is then concluded provided we show that

$$\inf_{\varphi\,\in\,\mathcal{D}(\Omega;\mathcal{C}^\infty(\mathbf{T}))}\int_\Omega h_\varphi\,dx = \int_\Omega\left\{\inf_{\varphi\,\in\,\mathcal{D}(\Omega;\mathcal{C}^\infty(\mathbf{T}))} h_\varphi(x)\right\}\,dx\ . \tag{5.7}$$

At this point we make use of the following commutation argument [8, Theorem 1]: if $\mathcal{H}\subset L^1(\Omega)$ is an *inf-stable family*, in the sense that

$$\begin{gathered}\{u_1,\dots,u_N\}\subset\mathcal{H}\ ,\quad \{\alpha_1,\dots,\alpha_N\}\subset\mathcal{C}^1(\overline{\Omega};[0,1]),\ \text{with}\ \sum_{i=1}^N\alpha_i = 1,\\ \Downarrow \\ \exists u\in\mathcal{H}\ :\ u\le\sum_{i=1}^N\alpha_i u_i\ ,\end{gathered} \tag{5.8}$$

then

$$\inf_{u\in\mathcal{H}}\int_\Omega u\,dx=\int_\Omega \operatorname*{ess\,inf}_{u\in\mathcal{H}} u\;dx\;.$$

In order to prove (5.7), we apply this principle to the the subfamily of $L^1(\Omega)$ defined by

$$\mathcal{H}:=\big\{h_\varphi\;:\;\varphi\in\mathcal{D}(\Omega;\mathcal{C}^\infty(\mathbf{T}))\big\}\;.$$

To check the inf-stability of $\mathcal{H}$, we consider $\{\varphi_1,\dots,\varphi_N\}\subset\mathcal{D}(\Omega;\mathcal{C}^\infty(\mathbf{T}))$, $\{\alpha_1,\dots,\alpha_N\}$ as in (5.8), and $\overline{\varphi}(x,y):=\sum_{i=1}^N\alpha_i(x)\varphi(x,y)$. Then $\overline{\varphi}\in\mathcal{D}(\Omega;\mathcal{C}^\infty(\mathbf{T}))$, so that $h_{\overline{\varphi}}\in\mathcal{H}$. We claim that $h_{\overline{\varphi}}\le\sum_{i=1}^N\alpha_i h_{\varphi_i}$. Indeed,

$$\begin{aligned}h_{\overline{\varphi}}(x)&:=\int_Y j\big(y,\nabla u(x)+\nabla_y\overline{\varphi}(x,y)\big)d\mu(y)\\&=\int_Y j\left(y,\sum_{i=1}^N\alpha_i(x)[\nabla u(x)+\nabla_y\varphi_i(x,y)]\right)d\mu(y)\\&\le\sum_{i=1}^N\alpha_i(x)\int_Y j\big(y,\nabla u(x)+\nabla_y\varphi_i(x,y)\big)\,d\mu(y)=\sum_{i=1}^N\alpha_i(x)h_{\varphi_i}(x),\end{aligned}$$

where we used the convexity of j in its second variable. Thus we get

$$\inf_{\varphi\,\in\,\mathcal{D}(\Omega;\mathcal{C}^\infty(\mathbf{T}))}\int_\Omega h_\varphi(x)\,dx=\int_\Omega\operatorname*{ess\,inf}_{\varphi\,\in\,\mathcal{D}(\Omega;\mathcal{C}^\infty(\mathbf{T}))}h_\varphi(x)\,dx\;.\tag{5.9}$$

On the other hand, $\mathcal{H}$ is contained in the space $\mathcal{C}_0(\Omega)$, which is separable with respect to the uniform norm; hence, we can find a sequence $\{h_n\}$ such that $\inf_{h\in\mathcal{H}}h(x)=\inf_n h_n(x)\;\forall x\in\Omega$. It follows from the definition of essential infimum that

$$\operatorname*{ess\,inf}_{h\in\mathcal{H}}h(x)=\inf_n h_n(x)=\inf_{h\in\mathcal{H}}h(x)\qquad\mathcal{L}^n\text{-a.e. on }\Omega\;.\tag{5.10}$$

By (5.9) and (5.10), (5.7) is proved and this achieves our proof. □

As a consequence of well-known properties of Γ-convergence [19, Theorem 7.8 and Corollary 7.17], we deduce from Lemma 6.1 and Theorem 6.2 the following.

COROLLARY 5.4. *Let μ satisfy $(H4)$ with $p>1$, and consider for any $\varepsilon>0$ the problem*

$$\inf\left\{J_\varepsilon(u)-\int_\Omega fu\,d\mu_\varepsilon\;,\quad u\in\mathcal{C}_0^1(\Omega)\right\},\tag{$\mathcal{P}_\varepsilon$}$$

where J_ε are defined by (5.1)*, and f is a prescribed function in $\mathcal{C}(\overline{\Omega})$. Then*

(i) $\liminf_{\varepsilon\to0}(\mathcal{P}_\varepsilon)=\min(\mathcal{P})$*;*

(ii) *if $\{u_\varepsilon\}$ is a minimizing sequence for $(\mathcal{P}_\varepsilon)$, up to subsequences it holds, that $u_\varepsilon\mu_\varepsilon\rightharpoonup u\mathcal{L}^n$, where u solves the limit problem $(\mathcal{P})$, given by*

$$\min\left\{J^{\rm hom}(u)-\int_\Omega fu\,dx\;,\quad u\in W_0^{1,p}(\Omega)\right\}.\tag{$\mathcal{P}$}$$

As usual, in the case of quadratic functionals, the unit-cell problem defined in (5.2) is completely determined in terms of n linear problems. Let us assume that the density of J_ε is given by $j(y,z) = \frac{1}{2}\sum_{i,j} a_{ij}(y) z_i\, z_j$, where the coefficients of the matrix $A(y) := a_{ij}(y)$ belong to $L^\infty(\mathbf{T})$ and are such that $\lambda|z|^2 \le j(y,z) \le \Lambda(1+|z|^2)$ for suitable positive constants λ and Λ. Then, we consider the following linear problems on the unit cell:

$$-\operatorname{div}_y \Big[A(y)\Big(e_{i,\mu}(y) + \nabla_\mu \chi_i(y)\Big)\Big] = 0\ , \qquad \chi_i \in H^{1,2}_\mu(\mathbf{T})\ , \qquad i = 1,2,\dots,n\ , \tag{5.11}$$

where we have set $e_{i,\mu}(y) := P_\mu(y) e_i$. Accordingly, the homogenized integrand takes the form $j^{\hom}(z) = \frac{1}{2} A^{\hom} z \cdot z$, where the effective matrix $A^{\hom}$ is given by

$$A^{\hom}_{i,j} = \int_Y \big(e_{i,\mu} + \nabla_{\mu,y}\chi_i\big)\cdot\big(e_{j,\mu} + \nabla_{\mu,y}\chi_j\big)\, d\mu(y)\ ; \tag{5.12}$$

consequently, the homogenized problem reads as

$$-\operatorname{div}_x\big(A^{\hom}\nabla u(x)\big) = f(x)\ , \qquad u \in W^{1,2}_0(\Omega). \tag{5.13}$$

In light of our two-scale approach we obtain the following corrector-type result.

PROPOSITION 5.5. *Under the same assumptions of Corollary* 5.4, *suppose in addition that j is quadratic as above and let $\{u_\varepsilon\}$ be a minimizing sequence for problem $(\mathcal{P}_\varepsilon)$. Then*

$$(u_\varepsilon, \nabla_{\mu_\varepsilon} u_\varepsilon) \ \rightharpoonup\ \left(u(x)\,,\ \sum_i (e_i + \nabla_{\mu,y}\chi_i(y))\,\frac{\partial u}{\partial x_i}(x)\right), \tag{5.14}$$

where $u \in W^{1,2}_0(\Omega)$ and $\chi_i \in H^{1,2}_\mu(\mathbf{T})$ are, respectively, the unique solutions to (5.13) *and to* (5.11). *Moreover, setting $u_1(x,y) := \sum_i \frac{\partial u}{\partial x_i}(x)\,\chi_i(y)$, we have*

$$\lim_{\varepsilon\to 0}\int_\Omega \big|u_\varepsilon(x) - u(x)\big|^2\, d\mu_\varepsilon = 0, \tag{5.15}$$

$$\lim_{\varepsilon\to 0}\Big\|u_\varepsilon - u - u_1\Big(x, \frac{x}{\varepsilon}\Big)\Big\|_{H^{1,2}_{\mu_\varepsilon}(\Omega)} = 0. \tag{5.16}$$

Proof. By Theorem 4.2, up to subsequences $(u_\varepsilon, \nabla_{\mu_\varepsilon} u_\varepsilon)$ two-scale converges to $(u(x), \nabla u(x) + \nabla_{\mu,y} u_1(x,y))$, where $(u,u_1) \in W^{1,2}_0(\Omega)\times L^2(\Omega, H^{1,2}_\mu(\mathbf{T}))$. Let now $(\varphi,\varphi_1) \in \mathcal{D}(\Omega)\times\mathcal{D}(\Omega;\mathcal{C}^\infty(\mathbf{T}))$, and multiply by $\varphi(x) + \varphi_1(x, \frac{x}{\varepsilon})$ the Euler equation satisfied by u_ε. Passing to the two-scale limit, we obtain that (u,u_1) satisfy the variational formulation

$$\begin{aligned}&\int_{\Omega\times Y} A(y)\,[\nabla u(x) + \nabla_{\mu,y} u_1(x,y)]\cdot[\nabla\varphi(x) + \nabla_{\mu,y}\varphi_1(x,y)]\, dm\\ &\qquad = \int_\Omega f(x)\varphi(x)\, dx \qquad \forall(\varphi,\varphi_1)\in\mathcal{D}(\Omega)\times\mathcal{D}(\Omega;\mathcal{C}^\infty(\mathbf{T}))\ .\end{aligned} \tag{5.17}$$

By Lax–Milgram's theorem, (5.17) admits a unique solution; thus (u,u_1) do not depend on the choice of the subsequence. Now it is easily seen that (5.17) is the variational formulation of the system

$$\begin{cases} -\operatorname{div}_y\Big[A(y)\Big(\nabla u(x) + \nabla_{\mu,y} u_1(x,y)\Big)\Big] = 0 & \text{in } \Omega\times Y,\\ -\operatorname{div}_x\Big[\displaystyle\int_Y A(y)\Big(\nabla u(x) + \nabla_{\mu,y} u_1(x,y)\Big)\, d\mu(y)\Big] = f & \text{in } \Omega,\end{cases}$$

which is in turn equivalent to the homogenized equations (5.13) and (5.11) via (5.12) and the relation $u_1(x,y) = \sum_i \frac{\partial u}{\partial x_i}(x)\,\chi_i(y)$. Thus (5.14) is proved. In order to prove the second part of the statement, we notice first that the corrector result (5.15) follows straightforwardly from (5.16). Indeed, by (H4), a Poincaré constant independent of ε rules as well for all the measures μ_ε (cf. (5.5)). Hence, we are done if we show (5.16). First we claim that, since u_ε is a minimizing sequence for $(\mathcal{P}_\varepsilon)$, we have

$$\lim_{\varepsilon\to 0}\left(\int_\Omega A\left(\frac{x}{\varepsilon}\right)\left\{\nabla_{\mu_\varepsilon} u_\varepsilon\right\}^2 dx - \int_\Omega f u_\varepsilon dx\right) = 0\ . \tag{5.18}$$

Indeed, setting $\alpha_\varepsilon := \int_\Omega A(\frac{x}{\varepsilon})\big\{\nabla_{\mu_\varepsilon} u_\varepsilon\big\}^2 dx$ and $\beta_\varepsilon := \int_\Omega f u_\varepsilon dx$, we have, for every real t and for a suitable sequence $\delta_\varepsilon \to 0$,

$$\frac{1}{2}\alpha_\varepsilon - \beta_\varepsilon = J_\varepsilon(u_\varepsilon) - \int_\Omega f u_\varepsilon\, d\mu_\varepsilon \;\le\; J_\varepsilon(tu_\varepsilon) - \int_\Omega f\,(tu_\varepsilon)\, d\mu_\varepsilon + \delta_\varepsilon = \frac{1}{2}t^2\alpha_e - t\beta_\varepsilon + \delta_\varepsilon\ .$$

Taking then the infimum with respect to t, we get $\frac{1}{2}\alpha_\varepsilon - \beta_\varepsilon \le -\frac{1}{2}\frac{\beta_\varepsilon^2}{\alpha_\varepsilon} + \delta_\varepsilon$; thus $\frac{(\alpha_\varepsilon-\beta_\varepsilon)^2}{\alpha_e} \to 0$, which proves the claim. Now, by (5.18),

$$\begin{aligned}
&\lim_{\varepsilon\to 0}\int_\Omega A\left(\frac{x}{\varepsilon}\right)\left\{\nabla_{\mu_\varepsilon}\left[u_\varepsilon(x) - u(x) - \varepsilon u_1\left(x,\frac{x}{\varepsilon}\right)\right]\right\}^2 d\mu_\varepsilon\\
&= \lim_{\varepsilon\to 0}\left\{\int_\Omega f(x)u_\varepsilon(x)d\mu_\varepsilon - 2\int_\Omega A\left(\frac{x}{\varepsilon}\right)\nabla_{\mu_\varepsilon}u_\varepsilon\cdot\nabla_{(\mu_\varepsilon)}\left[u(x)+\varepsilon u_1\left(x,\frac{x}{\varepsilon}\right)\right] d\mu_\varepsilon\right.\\
&\qquad\left. + \int_\Omega A\left(\frac{x}{\varepsilon}\right)\left\{\nabla_{\mu_\varepsilon}\left[u(x)+\varepsilon u_1\left(x,\frac{x}{\varepsilon}\right)\right]\right\}^2 d\mu_\varepsilon\right\}\\
&= \int_\Omega f(x)u(x)\,dx - \int_{\Omega\times Y} A(y)\left\{\nabla u(x) + \nabla_{\mu,y}u_1(x,y)\right\}^2 dm,
\end{aligned}$$

where we used the notation $Av^2 := Av\cdot v$. We now observe that, in view of the equality $u_1(x,y) = \sum_i \frac{\partial u}{\partial x_i}(x)\,\chi_i(y)$, it turns out that u_1, $\nabla_x u_1$, and $\nabla_{\mu,y}u_1$ are all "admissible" in the two-scale convergence (cf. Remark 2.4). Hence, we are allowed to take $(\varphi,\varphi_1) = (u,u_1)$ as test functions in (5.17); thus the last term in the above chain of equalities is zero, and (5.16) follows straightforward using the coercivity of A. □

Example 5.6. We conclude this section with an example of explicit computation of $j^{\hom}$ when μ is, respectively, the one-dimensional measure on a periodic grid or it has a Cantor-like structure. In the former case, our method allows us to recover in a very direct and concise way the homogenized energy density obtained in [17] through the classical fattening approach. The latter case can also be easily handled, in spite of the irregular structure of μ.

Let j be a quadratic-like energy density of the kind

$$j(z) = \frac{1}{2}\big(a_{11}z_1^2 + 2a_{12}z_1z_2 + a_{22}z_2^2\big)\ ;$$

we suppose the coefficients a_{ij} are constant and satisfy the nondegenerate ellipticity condition of Proposition 5.5 above (which, in particular, applies).

To the aim of computing $j^{\hom}$, we make use of Lemma 4.5(iii) so that we need the expression of j_μ for the measures μ under consideration. In view of the definitions of μ

specified below, only the coordinate axes e_1 and e_2 are involved as tangent directions. When $T_\mu(y) = \langle e_1 \rangle$, using definition (3.13) we find

$$j_\mu(y,z) = \inf_{\xi_2 \in \mathbb{R}} \frac{1}{2}\big(a_{11}z_1^2 + 2a_{12}z_1\xi_2 + a_{22}\xi_2^2\big) = \frac{1}{2}\left(a_{11} - \frac{a_{12}^2}{a_{22}}\right) z_1^2 =: \frac{1}{2} q_1 z_1^2.$$

Similarly, when $T_\mu(y) = \langle e_2 \rangle$, we obtain

$$j_\mu(y,z) = \frac{1}{2}\Big(a_{22} - \frac{a_{12}^2}{a_{11}}\Big) z_2^2 =: \frac{1}{2} q_2 z_2^2 \ .$$

Note that, by the uniform coercivity of the matrix (a_{ij}), the coefficients $q_1 = \frac{\det a_{i,j}}{a_{2,2}}$ and $q_2 = \frac{\det a_{i,j}}{a_{1,1}}$ are both strictly positive.

(i) Let μ be given on the unit cell $[-\frac{1}{2}, \frac{1}{2}[^2$ by $\mu := \frac{1}{2}\,(dx_1 \otimes \delta_0(dx_2) + \delta_0(dx_1) \otimes dx_2)$, δ_0 being the Dirac mass at 0. We obtain

$$\begin{aligned} j^{\hom}(z) &= \frac{1}{4} \min \left\{ \int_{-\frac{1}{2}}^{\frac{1}{2}} q_1 \big(z_1 + \nabla_1 u\,(y_1, 0)\big)^2 \, dy_1 \right. \\ &\qquad \left. + \int_{-\frac{1}{2}}^{\frac{1}{2}} q_2 \big(z_2 + \nabla_2 u\,(0, y_2)\big)^2 \, dy_2 \ : \ u \in H^{1,p}_\mu(\mathbf{T}) \right\} \\ &= \frac{1}{4}\big\{q_1 z_1^2 + q_2 z_2^2\big\}. \end{aligned}$$

(ii) Let μ be given on the unit cell $[-\frac{1}{2}, \frac{1}{2}[^2$ by $\mu := \frac{1}{2}\,(dx_1 \otimes \delta_0(dx_2) + \tau(dx_1) \otimes dx_2)$, τ being a probability measure concentrated on a Cantor subset of $(-\frac{1}{2}, \frac{1}{2})$ whose tangent space degenerates to $\{0\}$. Noticing that the condition $u \in H^{1,p}_\mu(\mathbf{T})$ implies that $\lim_{y_2 \to 0} u(y_1, y_2) = u(y_1, 0)$ for τ-a.e. y_1, it is easy to check that, as in case (i), μ satisfies assumption $(H4)$. However, the transmission condition along the segment $Y \cap \{y_1 = 0\}$ does not affect the value of the minimum problem involved in the definition of $j^{\hom}$. We actually obtain

$$\begin{aligned} j^{\hom}(z) &= \frac{1}{4} \min \left\{ \int_{-\frac{1}{2}}^{\frac{1}{2}} q_1 \big(z_1 + \nabla_1 u\,(y_1, 0)\big)^2 \, dy_1 \right. \\ &\qquad \left. + \int_{-\frac{1}{2}}^{\frac{1}{2}} \left(\int_{-\frac{1}{2}}^{\frac{1}{2}} q_2 \big(z_2 + \nabla_2 u\,(y_1, y_2)\big)^2 \, dy_2 \right) \tau(dy_1) \ : \ u \in H^{1,p}_\mu(\mathbf{T}) \right\} \\ &= \frac{1}{4}\big\{q_1 z_1^2 + q_2 z_2^2\big\}. \end{aligned}$$

6. The case of two-parameter integrals. Our framework of homogenization with periodic measures includes in a natural way the case of thin reinforced structures: they are made of identical cells periodically repeated, in which the material is concentrated along bars of thickness δ, small with respect to the period ε. Thus, the energy of the δ-thick, ε-periodic structure is given by

$$J_{\delta,\varepsilon}(u) = \int_\Omega j\big(\nabla u\big)\, d\mu_{\delta,\varepsilon} \ ; \tag{6.1}$$

here $\mu_{\delta,\varepsilon}$ is the ε-periodization of μ_δ according to (2.1), μ_δ being the Y-periodic measure associated with the structure of thickness δ. Notice that here we use a

simplified model where the density of energy in (6.1) is described by a homogeneous integrand $j(z)$ (in full generality $j(z)$ should be replaced by $j_\delta(\frac{x}{\varepsilon}, z)$).

In this section, we investigate the commutativity of the passage to the limit in (6.1) as the two parameters ε and δ tend to zero. In fact, a possible first procedure is to homogenize with respect to each μ_δ, and then let δ tend to zero; the computations required for the latter limit process can be found, in some particular cases, in [17], [15] (see also [18] for a quite recent survey about reticulated structures). A second natural procedure is to apply Theorem 5.2 taking as a measure μ the weak limit of μ_δ as $\delta \to 0$, namely, the measure associated with the skeleton of the structure. Let us emphasize that the latter method is in practice much simpler to apply, as in many cases it is easy to compute explicitly the effective energy density given by (4.2). Therefore it is worth establishing whether the two procedures are equivalent or not. Our claim is that the equivalence holds provided μ is connected. Actually, under this assumption, we can prove a much stronger statement: if we let δ depend on ε in (6.1), the Γ-limit of $J_{\delta(\varepsilon),\varepsilon}$ is the same whatever the choice of the sequence $\delta(\varepsilon)$. This is derived from Theorem 5.2 when the measures μ_δ are approximations of μ by convolution (see Theorem 6.1). Similar conclusions can be reached also in the case of δ-fattened structures (see Remark 6.2); we omit the required additional computations which are technical but straightforward.

In what follows, for any $\lambda > 0$, we let ρ_λ be a convolution kernel $\rho_\lambda(x) := \frac{1}{\lambda^n}\rho\left(\frac{x}{\lambda}\right)$, where ρ is assumed to be a smooth, positive, even function, compactly supported on the unit ball of $\mathbb{R}^n$, and such that $\int_{\mathbb{R}^n} \rho\, dx = 1$. For any measure ν, the notation $\rho_\lambda \star \nu$ will be used to denote the smooth function $\rho_\lambda \star \nu(x) := \int_{\mathbb{R}^n} \rho_\lambda(x-y)\, d\nu(y)$; we also recall that, when ν is Y-periodic, ν_ε denotes the measure defined by (2.1). In view of these definitions, it is easy to check that the measure $[(\rho_\lambda \star \nu)\mathcal{L}^n]_\varepsilon$ agrees with the measure with Lebesgue density $\rho_{\lambda\varepsilon} \star \nu_\varepsilon$.

We can now give the main result of the section. We stress that the Γ-limit of $J_{\delta,\varepsilon}$ is computed with respect to the analogous convergence as in Theorem 5.2, namely, $u_\varepsilon \mu_{\delta,\varepsilon} \rightharpoonup u\mathcal{L}^n$.

THEOREM 6.1. *Let μ satisfy $(H4)$, with $p > 1$. Let ε and $\delta = \delta(\varepsilon)$ be positive parameters tending to zero, and let $\mu_{\delta,\varepsilon} := [(\rho_{\delta(\varepsilon)} \star \mu)\mathcal{L}^n]_\varepsilon$. Then the functionals $J_{\delta,\varepsilon}$ defined by* (6.1) *Γ-converge to the functional $J^{\hom}$ defined by* (5.6).

Proof. For simplicity, throughout the proof we shall not denote the dependence of δ on ε. We begin by checking the Γ-liminf inequality for any sequence u_ε such that $u_\varepsilon \mu_{\delta,\varepsilon} \rightharpoonup u\mathcal{L}^n$. If we set $v_\varepsilon = \rho_{\delta\cdot\varepsilon} \star u_\varepsilon$, by the Jensen's inequality we have

$$\begin{aligned} J_{\delta,\varepsilon}(u_\varepsilon) &= \int j(\nabla u_\varepsilon)[\rho_{\delta\cdot\varepsilon} \star \mu_\varepsilon]\, dx = \int \rho_{\delta\cdot\varepsilon} \star j(\nabla u_\varepsilon)\, d\mu_\varepsilon \\ &\geq \int j(\rho_{\delta\cdot\varepsilon} \star \nabla u_\varepsilon)\, d\mu_\varepsilon = J_\varepsilon(v_\varepsilon). \end{aligned}$$

Hence, the Γ-liminf inequality is a straightforward consequence of Theorem 5.2, provided we prove that

$$v_\varepsilon \mu_\varepsilon \rightharpoonup u\mathcal{L}^n. \tag{6.2}$$

To show (6.2), let φ be any test function in $\mathcal{C}_0(\mathbb{R}^n)$. By Fubini's theorem, taking into account the assumption $\rho(z) = \rho(-z)$, we have

$$\begin{aligned}
&\left| \int \varphi \, (\rho_{\delta\cdot\varepsilon} \star u_\varepsilon) \, d\mu_\varepsilon - \int \varphi \, (\rho_{\delta\cdot\varepsilon} \star \mu_\varepsilon) \, u_\varepsilon \, d\mathcal{L}^n \right| \\
&\quad = \left| \int \int [\varphi(x) - \varphi(y)] \rho_{\delta\cdot\varepsilon}(x-y) \, u_\varepsilon(y) \, dy \, d\mu_\varepsilon(x) \right| \\
&\quad \leq \sup_{|s-t|<\delta\cdot\varepsilon} |\varphi(s) - \varphi(t)| \int \int |u_\varepsilon(y| \, \rho_{\delta\cdot\varepsilon}(x-y) \, d\mu_\varepsilon(x) \\
&\quad = \mathrm{o}(\delta\cdot\varepsilon) \int |u_\varepsilon| \, d\mu_{\delta,\varepsilon}.
\end{aligned}$$

Hence, $\{v_\varepsilon \mu_\varepsilon\}$ is bounded in the space of measures and actually converges weakly, as $\varepsilon \to 0$, to the same limit as the sequence $\{u_\varepsilon \mu_{\delta,\varepsilon}\}$. The claim (6.2) follows.

It remains to prove the Γ- limsup inequality. To this aim, it is enough to consider the same sequence $\{u_\varepsilon\} \subset \mathcal{D}(\Omega)$ already employed in the proof of Theorem 5.2, namely

$$u_\varepsilon(x) = u(x) + \varepsilon\varphi\left(x, \frac{x}{\varepsilon}\right)$$

for an arbitrary function $\varphi \in \mathcal{D}(\Omega; \mathcal{C}^\infty(\mathbf{T}))$. To verify that such a sequence gives the required upper bound for the Γ- limsup of $J_{\delta,\varepsilon}$, it is enough to replace μ_ε by $\mu_{\delta,\varepsilon}$ in the proof of Theorem 5.2 part (II) (accordingly, it has to be noticed that the convergence (2.3) still holds if in its left-hand member we substitute μ_ε with $\mu_{\delta,\varepsilon}$). □

Remark 6.2. A variant of Theorem 6.1 can be obtained in the case of δ-fattened stuctures. Let us consider a multijunction body $S = S_1 \cup S_2 \cup \cdots \cup S_m$, where S_k is a k-dimensional Y-periodic subset of $\mathbb{R}^n$. We assume in addition that every S_k is piecewise C^2 and satisfies $\mathcal{H}^k(S_k \cap \partial Y) = 0$, so that the normalized measures $\mu_k := \left[\mathcal{H}^k(Y \cap S_k)\right]^{-1} \mathcal{H}^k \llcorner S_k$ satisfy $\mu_k(Y) = 1$. The δ-fattened structure can be written as $S_\delta = \cup_k S_{k,\delta}$ where $S_{k,\delta} := \{x \in \mathbb{R}^n : \mathrm{dist}(x, S_k) \leq \delta\}$. Then we represent the energy on S_δ through (6.1), where $\mu_\delta := \sum_k c_k \, \mu_{k,\delta}$, being $\mu_{k,\delta} := \left[\mathcal{L}^n(Y \cap S_{k,\delta})\right]^{-1} \mathcal{L}^n \llcorner S_{k,\delta}$ and c_k is a weight on each part of the multijunction. We notice that μ_δ has a piecewise constant Lebesgue density in $\mathbb{R}^n$ and one checks easily, by using the regularity assumption on each S_k, that $\{\mu_\delta\}$ converges weakly to $\mu := \sum_k c_k \mu_k$ as δ tends to zero. The expected limit density of energy can be represented by (4.2) using this definition of measure μ. Indeed, if such μ satisfies $(H4)$ and if $J_{\delta,\varepsilon}$ is given by (6.1) with μ_δ defined above, then there holds $J_{\delta(\varepsilon),\varepsilon} \xrightarrow{\Gamma} J^{\mathrm{hom}}$ for any sequences ε and $\delta = \delta(\varepsilon)$ tending to zero. In particular, the limit processes with respect to the parameters ε and δ are actually commutative.

We conclude by emphasizing that Theorem 6.1 does not apply if μ does not enjoy the connectedness property (H1). This is illustrated by the example below, where the case of a fiber reinforced structure is considered and the exponent p represents the growth of the energy under consideration. Our method applies if and only if the choice of p ensures the connectedness of μ. In the other cases, it has been recently proved by Bellieud and Bouchitté [4] that the homogenized energy obtained as a contemporary limit, as $\varepsilon \to 0$ and $\delta = \delta(\varepsilon) \to 0$, of $J_{\delta,\varepsilon}$ may contain a nonlocal term which depends on the velocity of convergence to zero of $\delta(\varepsilon)$.

Example 6.3. Consider the sequence μ_δ of Y-periodic measures given on Y by the Lebesgue measure with a density a_δ equal, respectively, to 1 on $Y \setminus F_\delta$ and to $\frac{k}{\delta^2}$ on F_δ, being F_δ a system of cylindrical fibers of very small radius δ, parallel to the three coordinated axes (see Figure 7.1A).

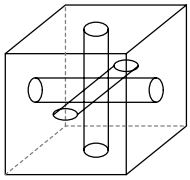


FIG. 7.1A.

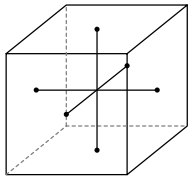


FIG. 7.1B.

Then $\mu_\delta \rightharpoonup \mu$, where $\mu = \mu' + \mu''$, being the restrictions of μ' and μ'' on Y given, respectively, by the Lebesgue measure $\mathcal{L}^3 \llcorner Y$ and by $k\pi$ times the one-dimensional measure $\mathcal{H}^1$ over the skeleton F of the fibers (see Figure 7.1B).

When $p > 2$, μ turns out to satisfy $(H4)$, so applying Theorem 5.2 with the choice $j(y,z) = \frac{1}{p}|z|^p$, we obtain that the Γ-limit of $J_{\delta,\varepsilon}$ is given by

$$J^{\rm hom}(u) := \frac{1}{p}\left\{\int_\Omega |\nabla u|^p\,dx + k\pi\int_\Omega \Big[\Big|\frac{\partial u}{\partial x_1}\Big|^p + \Big|\frac{\partial u}{\partial x_2}\Big|^p + \Big|\frac{\partial u}{\partial x_3}\Big|^p\Big]\,dx\right\},\quad u \in W^{1,p}_0(\Omega). \tag{6.3}$$

On the other hand, when $p \le 2$, μ fails to satisfy $(H1)$, so Theorem 4.2 does not apply, and the two-scale limit of a sequence $\{u_\varepsilon\}$ uniformly bounded in $L^p_{\mu_\varepsilon}(\Omega)$ takes a priori two different values on the two "connected components" of μ, i.e.

$$u_0(x,y) = \begin{cases} u(x) & \text{if } y \in Y \setminus F,\\ v(x) & \text{if } y \in F.\end{cases}$$

This kind of behavior suggests that one should consider the measures μ'_δ and μ''_δ given on Y, respectively, by $\mu_\delta \llcorner (Y \setminus F_\delta)$ and $\mu_\delta \llcorner F_\delta$, and to split the energies $J_{\delta,\varepsilon}$ into

$$J_{\delta,\varepsilon}(u) = \frac{1}{p}\int_\Omega |\nabla u|^p\,d(\mu'_\delta)_\varepsilon + \frac{1}{p}\int_\Omega |\nabla u|^p\,d(\mu''_\delta)_\varepsilon\ . \tag{6.4}$$

Since μ'_δ and μ''_δ weakly converge, respectively, to μ' and μ'', Theorem 5.2 can now be applied to each of the sequences on the right-hand side of (6.4). In this way, denoting by u and v the weak limits associated, respectively, with the sequences $\{u_\varepsilon\,\mu'_{\delta,\varepsilon}\}$ and $\{u_\varepsilon\,\mu''_{\delta,\varepsilon}\}$, we obtain the following lower bound:

$$\liminf_{\varepsilon\to 0} J_{\delta(\varepsilon),\varepsilon}(u_\varepsilon) \ge \frac{1}{p}\left\{\int_\Omega |\nabla u|^p\,dx + k\pi\int_\Omega \Big[\Big|\frac{\partial v}{\partial x_1}\Big|^p + \Big|\frac{\partial v}{\partial x_2}\Big|^p + \Big|\frac{\partial v}{\partial x_3}\Big|^p\Big]\,dx\right\}.$$

Assuming that v agrees with u would lead us to replace the right-hand member of the above inequality by the functional $J^{\rm hom}(u)$ given by (6.3), which in fact does not provide the optimal lower bound. Indeed, the crucial point is that the correct limit energy $\Phi(u,v)$ contains an additional term (depending on the gap $v-u$) which takes into account the interaction between the two "connected components" of μ. Precisely, we have (see [4])

$$\Phi(u,v) = \frac{1}{p}\left\{\int_\Omega |\nabla u|^p\,dx + k\pi\int_\Omega \Big[\Big|\frac{\partial v}{\partial x_1}\Big|^p + \Big|\frac{\partial v}{\partial x_2}\Big|^p + \Big|\frac{\partial v}{\partial x_3}\Big|^p\Big]\,dx + 6\pi\gamma\int_\Omega |u-v|^p\,dx\right\},$$

where the positive constant γ is given by

$$\gamma = \begin{cases} \lim\limits_{\varepsilon\to 0}\ \varepsilon^{-2}\big|\log\big(\varepsilon\delta(\varepsilon)\big)\big|^{-1} & \text{if } p = 2,\\ \lim\limits_{\varepsilon\to 0}\ \Big|\dfrac{2-p}{p-1}\Big|^{p-1}\varepsilon^{-p}\delta(\varepsilon)^{2-p} & \text{if } p < 2\end{cases}$$

(notice that γ does depend on the way $\delta(\varepsilon)$ tends to zero). Thus the Γ-limit of $J_{\delta,\varepsilon}$ is given by

$$J(u) := \inf\left\{\Phi(u,v) \ : \ v \in W_0^{1,p}(\Omega)\right\}, \qquad u \in W_0^{1,p}(\Omega),$$

which means to compute the infimum with respect to v of a nonlocal energy in u (whatever γ is in $(0,+\infty)$).

We address to a forthcoming paper for a systematic treatment, in our framework of measures, of nonlocal phenomenon due to the lack of connectedness.

Acknowledgment. The second author wishes to thank the hospitality and support of the University of Toulon.

REFERENCES

[1] G. Allaire, *Homogenization and two-scale convergence*, SIAM J. Math. Anal., 23 (1992), pp. 1482–1518.

[2] G. Allaire, *Two-scale convergence: A new method in periodic homogenization*, in Nonlinear Partial Differential Equations and Their Applications, Collège de France Seminar, Vol. 12 (Paris, 1991–1993), Pitman Res. Notes Math. Ser. 302, Longman, Harlow, UK, 1994, pp. 1–14.

[3] H. Attouch and G. Buttazzo, *Homogenization of reinforced periodic one-codimensional structures*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 14 (1987), pp. 465–484.

[4] M. Bellieud and G. Bouchitté, *Homogenization of elliptic problems in a fiber reinforced structure. Non local effects*, Ann. Scuola Norm. Sup. Cl. Sci. (4), 26 (1998), pp. 407–436.

[5] A. Bensoussan, J.-L. Lions, and G. Papanicolaou, *Asymptotic Analysis for Periodic Structures*, North-Holland, Amsterdam, 1978.

[6] G. Bouchitté, G. Buttazzo, and I. Fragalà, *Mean curvature of a measure and related variational problems*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 25 (1997), pp. 179–196.

[7] G. Bouchitté, G. Buttazzo, and P. Seppecher, *Energies with respect to a measure and applications to low dimensional structures*, Calc. Var. Partial Differential Equations, 5 (1997), pp. 37–54.

[8] G. Bouchitté and M. Valadier, *Integral representation of convex functionals on a space of measures*, J. Funct. Anal., 80 (1988), pp. 398–420.

[9] A. Braides and A. Defranceschi, *Homogenization of Multiple Integrals*, Oxford Lecture Ser. Math. Appl. 12, Clarendon Press, Oxford, UK, 1998.

[10] H. Brézis, *Analyse Fonctionnelle*, Masson, Paris (1993).

[11] M. Briane, *Poincaré-Wirtinger's inequality for the homogenization in perforated domains*, Boll. Uni. Mat. Ital. B (7), 11 (1997), pp. 53–82.

[12] G. Buttazzo, *Semicontinuity, Relaxation, and Integral Representation in the Calculus of Variations*, Pitman Res. Notes Math. Ser. 207, Longman, Harlow, UK, 1989.

[13] C. Castaing and M. Valadier, *Convex Analysis and Measurable Multifunctions*, Lecture Notes in Math. 580, Springer-Verlag, Berlin, 1977.

[14] V. Chiadò Piat, G. Dal Maso, and A. Defranceschi, *G-convergence of monotone operators*, Ann. Inst. H. Poincaré, Anal. Non Linéaire, 7 (1990), pp. 123–160.

[15] R. Chiheb, D. Cioranescu, A. El Janati, and G. Panasenko, *Structures réticulées renforcées en elasticité*, C. R. Acad. Sci. Paris Sér. I Math., 326 (1998), pp. 879–902.

[16] D. Cioranescu and J. Saint Jean Paulin, *Homogenization in open sets with holes*, J. Math. Anal. Appl., 71 (1979), pp. 590–607.

[17] D. Cioranescu and J. Saint Jean Paulin, *Reinforced and honey-comb structures*, J. Math. Pures Appl. (9), 65 (1986), pp. 403–422.

[18] D. Cioranescu and J. Saint Jean Paulin, *Homogenization of Reticulated Structures*, Appl. Math. Sci. 136, Springer-Verlag, New York, 1999.

[19] G. Dal Maso, *An Introduction to* Γ*-convergence*, Birkhäuser, Boston, 1993.

[20] B. Dacorogna, *Direct Methods in the Calculus of Variations*, Appl. Math. Sci. 78, Springer-Verlag, Berlin, 1988.

[21] I. Fragalà and C. Mantegazza, *On some notions of tangent space to a measure*, Proc. Roy. Soc. Edinburgh, Sect. A 129 (1999), pp. 331–342.

[22] P. Hajlasz and P. Kosela, *Sobolev met Poincaré*, Mem. Amer. Math. Soc., 145 (2000), number 688.

[23] J.-L. Lions, *Some Methods in the Mathematical Analysis of Systems and Their Control*, Science Press, Beijing, Gordon Press, New York, 1981.
[24] P. Marcellini, *Periodic solutions and homogenization of non-linear variational problems*, Ann. Mat. Pura Appl., 117 (1978), pp. 139–152.
[25] F. Murat and L. Tartar, *H-convergence*, in Topics in the Mathematical Modelling of Composite Materials, Progr. Nonlinear Differential Equations Appl. 31, A. Cherkaev and R. V. Kohn, eds., Birkhäuser, Boston, 1997, pp. 21–43.
[26] G. Nguetseng, *A general convergence result for a functional related to the theory of homogenization*, SIAM J. Math. Anal., 20 (1989), pp. 608–623.
[27] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[28] V. V. Zhikov, *Connectedness and homogenization. Examples of fractal conductivity*, Mat. Sb., 187 (1996), pp. 1109–1147.

SIAM J. MATH. ANAL.
Vol. 32, No. 6, pp. 1227–1247

THE FINITE LARMOR RADIUS APPROXIMATION*

EMMANUEL FRÉNOD† AND ERIC SONNENDRÜCKER‡

Abstract. The presence of a large external magnetic field in a plasma introduces an additional time-scale which is very constraining for the numerical simulation. Hence it is very useful to introduce averaged models which remove this time-scale. However, depending on other parameters of the plasma, different starting models for the asymptotic analysis may be considered. We introduce here a generic framework for our analysis which fits many of the possible regimes and apply it in particular to justify the finite Larmor radius approximation both in the linear case and in the nonlinear case in the plane transverse to the magnetic field.

Key words. Vlasov–Poisson equations, kinetic equations, homogenization, two-scale convergence, multiple time scales

AMS subject classifications. 82D10, 35B27, 35Q99, 76X05

PII. S0036141099364243

1. Introduction.

The main goal of this paper is the investigation of an asymptotic regime taking place in the description of the behavior of charged particles under the action of a strong external magnetic field and called the finite Larmor radius approximation. This approximation has a natural field of application in tokamak physics.

This work was announced in Frénod and Sonnendrücker [9] and follows Frénod and Sonnendrücker [8, 10], where we exhibited global asymptotic behavior of plasmas. Those global behaviors have also been mathematically put in light by Golse and Saint-Raymond [12, 11] and Grenier [14]. The context of the finite Larmor radius approximation is more local. Its object is to describe the behavior of the considered plama's particles when the observation length scale is comparable with their Larmor radius.

We choose to lead our study in the framework of the Vlasov equation which writes, in this context

$$\begin{aligned} &\frac{\partial f^\epsilon}{\partial t} + \mathbf{v}_\parallel \cdot \nabla_x f^\epsilon + \frac{1}{\epsilon}\mathbf{v}_\perp \cdot \nabla_x f^\epsilon + \left(\mathbf{E} + \frac{1}{\epsilon}\mathbf{v}\times\mathbf{m}\right)\cdot \nabla_v f^\epsilon = 0, \\ &f^\epsilon_{|t=0} = f_0, \end{aligned} \tag{1.1}$$

where ϵ is a small parameter which will tend to 0. In (1.1) the distribution function $f^\epsilon \equiv f^\epsilon(t,\mathbf{x},\mathbf{v})$; $t \in [0,T)$ for some $T < \infty$ is the time, $\mathbf{x} = (x_1,x_2,x_3) \in \mathbb{R}^3_x$ is the position, and $\mathbf{v} = (v_1,v_2,v_3) \in \mathbb{R}^3_v$ is the velocity. We denote $\mathcal{O} = \mathbb{R}^3_x \times \mathbb{R}^3_v$, $\Omega = [0,T) \times \mathbb{R}^3_x$, and $\mathcal{Q} = [0,T) \times \mathcal{O}$. The magnetic field $\mathbf{m}$ is supposed to be $\mathbf{e}_1$, the first vector of the frame $(\mathbf{e}_1,\mathbf{e}_2,\mathbf{e}_3)$ of $\mathbb{R}^3$. For any vector $\mathbf{v} \in \mathbb{R}^3$, $\mathbf{v}_\parallel$ stands for $\mathbf{v}_\parallel = (\mathbf{v}\cdot\mathbf{m})\mathbf{m} = v_1\mathbf{e}_1$ and $\mathbf{v}_\perp$ for $\mathbf{v}_\perp = \mathbf{v} - \mathbf{v}_\parallel = v_2\mathbf{e}_2 + v_3\mathbf{e}_3$. The electric field $\mathbf{E} \equiv \mathbf{E}(t,\mathbf{x})$ is external and nonoscillating.

*Received by the editors November 15, 1999; accepted for publication (in revised form) September 15, 2000; published electronically February 28, 2001.
http://www.siam.org/journals/sima/32-6/36424.html

†Laboratoire de Mathématiques et Applications des Mathématiques, Université de Bretagne Sud, BP573, Campus de Tohannic, F-56017 Vannes, France (emmanuel.frenod@univ-ubs.fr).

‡CNRS, Institut Elie Cartan, Université Henri Poincaré, Nancy and Institut de Recherche en Mathématiques Avancées, Université Louis Pasteur, 7 rue Rene Descartes, F-67084 Strasbourg Cedex, France (sonnen@math.u-strasbg.fr).

In order to make the process $\epsilon \to 0$ in (1.1), we assume

$$f_0 \geq 0, \quad f_0 \in L^1 \cap L^2(\mathcal{O}), \tag{1.2}$$

and for $\mathbf{E}$, we assume

$$\mathbf{E} \in \mathcal{C}^1(\Omega). \tag{1.3}$$

Then we have the following theorem.

THEOREM 1.1. *Under assumptions* (1.2) *and* (1.3)*, for each* $\epsilon > 0$*, there exists a unique solution* f^ϵ *of* (1.1) *in* $L^\infty(0,T,L^1 \cap L^2(\mathcal{O}))$*. As* $\epsilon \to 0$*,*

$$f^\epsilon \rightharpoonup f \text{ in } L^\infty(0,T,L^2(\mathcal{O})) \text{ weak} - *, \tag{1.4}$$

where f *is the unique solution of*

(1.5)
$$\frac{\partial f}{\partial t} + \mathbf{v}_\parallel \cdot \nabla_x f + \frac{1}{2\pi}\left(\int_0^{2\pi} \mathcal{R}(-\tau)\mathbf{E}(t, \mathbf{x} + \mathcal{R}(\tau)\mathbf{v})\, d\tau\right) \cdot \nabla_x f + \frac{1}{2\pi}\left(\int_0^{2\pi} R(-\tau)\mathbf{E}(t, \mathbf{x} + \mathcal{R}(\tau)\mathbf{v})\, d\tau\right) \cdot \nabla_v f = 0,$$

$$f_{|t=0} = \frac{1}{2\pi}\int_0^{2\pi} f_0(\mathbf{x} + \mathcal{R}(-\tau)\mathbf{v}, R(-\tau)\mathbf{v})\, d\tau,$$

where the matrices R *and* $\mathcal{R}$ *are given by*

$$R(\tau) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\tau & \sin\tau \\ 0 & -\sin\tau & \cos\tau \end{pmatrix}, \quad \mathcal{R}(\tau) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \sin\tau & 1-\cos\tau \\ 0 & \cos\tau - 1 & \sin\tau \end{pmatrix}. \tag{1.6}$$

The way to prove this theorem uses the 2-scale convergence defined as follows.

THEOREM 1.2 (see Nguetseng [18] and Allaire [2]). *If a sequence* f^ϵ *is bounded in* $L^\infty(0,T;W)$*, for a Banach spaces* W *being the dual of a separable space and being compactly embedded in* $\mathcal{D}'(\mathcal{O})$*, then for every period* θ *there exists a* θ*-periodic profile* $F_\theta(t,\tau,\mathbf{x},\mathbf{v}) \in L^\infty(0,T;L^\infty_\theta(\mathbb{R}_\tau;W))$ *such that for all* $\psi_\theta(t,\tau,\mathbf{x},\mathbf{v})$ *regular, with compact support with respect to* $(t,\mathbf{x},\mathbf{v})$ *and* θ*-periodic with respect to* τ*, we have, up to a subsequence,*

$$\int_{\mathcal{Q}} f^\epsilon \psi^\epsilon_\theta\, dt\, d\mathbf{x}\, d\mathbf{v} \to \int_{\mathcal{Q}}\int_0^\theta F_\theta \psi_\theta\, d\tau\, dt\, d\mathbf{x}\, d\mathbf{v}. \tag{1.7}$$

We then say that f^ϵ *two scale converges to* F_θ*. Above,* $L^\infty_\theta(\mathbb{R}_\tau)$ *stands for the space of functions being* $L^\infty(\mathbb{R})$ *and being* θ*-periodic and* $\psi^\epsilon_\theta \equiv \psi_\theta(t, \frac{t}{\epsilon}, \mathbf{x}, \mathbf{v})$*.*

The profile F_θ *is called the* θ*-periodic two scale limit of* f^ϵ *and the link between* F_θ *and the weak–*$*$ *limit* f *of* (f^ϵ) *is given by*

$$\int_0^\theta F_\theta(t,\tau,\mathbf{x},\mathbf{v})\, d\tau = f(t,\mathbf{x},\mathbf{v}). \tag{1.8}$$

Moreover, if a sequence (g^ϵ) *strongly converges to* g *in a second Banach space* W' *(with the same assumption for* W' *as for* W*), such that the product makes sense in a third Banach space* W''*, then,*

$$f^\epsilon g^\epsilon \text{ 2-scale converges to } F_\theta g \in L^\infty(0,T;L^\infty_\theta(\mathbb{R}_\tau;W'')). \tag{1.9}$$

Remark. Our definition of the two scale convergence by (1.7) does not comply with the averaging rule usually used. Otherwise the right-hand side of (1.7) would be divided by θ.

The proof of Theorem 1.1 consists in finding a constraint equation for the two scale limit F of f^ϵ, using a weak formulation with oscillating test function of (1.1). This constraint imposes a form to F. Then using oscillating test functions satisfying the constraint equation in the previously evoked weak formulation gives the equation satisfied by F. Integrating this last equation yields finally (1.5).

As the proof of Theorem 1.1 in this paper and of Theorems 1.1 and 3.2 of Frénod and Sonnendrücker [8] are very close, we develop here a generic framework inside which all those proofs may be included. This generic framework consists in considering a conservation law linearly perturbed:

$$\begin{aligned} &\frac{\partial u^\epsilon}{\partial t} + \mathbf{A}\cdot\nabla_x u^\epsilon + \frac{1}{\epsilon}\mathbf{L}\cdot\nabla_x u^\epsilon = 0,\\ &u^\epsilon_{t=0} = u_0. \end{aligned} \tag{1.10}$$

In this system, $u^\epsilon \equiv u^\epsilon(t,\mathbf{x})$, $t\in[0,T)$ for some $T<\infty$ and $\mathbf{x}\in\mathbb{R}^n=\mathcal{O}$. Let us mention that $\mathbf{x}$ here is an abstract variable which is not connected to the position which is also denoted by $\mathbf{x}$ in the Vlasov equation. We denote $\mathcal{Q}=[0,T)\times\mathcal{O}$, and we assume $\mathbf{A}\equiv\mathbf{A}(t,\mathbf{x})\in L^\infty(0,T;L^2_{loc}(\mathcal{O}))$, with $\nabla_x\cdot\mathbf{A}=0$ and $\mathbf{L}\equiv M\mathbf{x}+N$, where M is a real $n\times n$ matrix with constant entries, satisfying $\mathrm{tr}M=0$ and where $N\in\mathrm{Im}M$. We moreover assume that $e^{\tau M}$ is θ-periodic for a given $\theta\in\mathbb{R}$. The generic theorem writes as follows.

THEOREM 1.3. *Under the assumptions above, if, moreover, the sequence (u^ϵ) of solution of* (1.10) *satisfies*

$$\|u^\epsilon\|_{L^\infty(0,T;L^2(\mathcal{O}))}\le C, \tag{1.11}$$

for some constants C independent on ϵ, then, extracting a subsequence,

$$u^\epsilon \text{ 2-scale converges to a } \theta\text{-periodic profile } U\in L^\infty(0,T;L^\infty_\theta(\mathbb{R}_\tau;L^2(\mathcal{O}))) \tag{1.12}$$

and

$$u^\epsilon \rightharpoonup u \text{ in } L^\infty(0,T;L^2(\mathcal{O})) \text{ weak}-*. \tag{1.13}$$

We have

$$U(t,\tau,\mathbf{x}) = U_0(t, e^{-\tau M}(\mathbf{x}-\overline{N})+\overline{N}), \tag{1.14}$$

where $\overline{N}$ is such that $-M\overline{N}=N$ and where $U_0\equiv U_0(t,\mathbf{y})$ is solution of

$$\begin{aligned} &\frac{\partial U_0}{\partial t} + \frac{1}{\theta}\int_0^\theta e^{-\sigma M}\mathbf{A}(t, e^{\sigma M}(\mathbf{y}-\overline{N})+\overline{N})\,d\sigma\cdot\nabla_y U_0 = 0,\\ &U_{0|t=0} = \tfrac{1}{\theta}u_0. \end{aligned} \tag{1.15}$$

Moreover, u is solution of

$$\begin{aligned} &\frac{\partial u}{\partial t} + \frac{1}{\theta}\int_0^\theta e^{-\sigma M}\mathbf{A}(t, e^{\sigma M}(\mathbf{x}-\overline{N})+\overline{N})d\sigma\cdot\nabla_x\, u = 0,\\ &u_{|t=0}(\mathbf{x}) = \frac{1}{\theta}\int_0^\theta u_0(e^{-\sigma M}(\mathbf{x}-\overline{N})+\overline{N})\,d\sigma. \end{aligned} \tag{1.16}$$

When restricting to the plane perpendicular to $\mathbf{m}$, we may extend the previous result to the Vlasov–Poisson system.

We suppose now that f^ϵ does not depend on x_1 and v_1, and we use the following notations: $t \in [0,T)$, $T < \infty$, still denotes the time, the position- and velocity-variables become $\mathbf{x} = (x_2, x_3) \in \mathbb{R}^2_x$ and $\mathbf{v} = (v_2, v_3) \in \mathbb{R}^2_v$. We set $\mathcal{O} = \mathbb{R}^2_x \times \mathbb{R}^2_v$, $\Omega = [0,T) \times \mathbb{R}^2_x$, and $\mathcal{Q} = [0,T) \times \mathcal{O}$. For clarity, we denote $\mathcal{O}' = \mathbb{R}^2_y \times \mathbb{R}^2_u$ and $\mathcal{Q} = [0,T) \times \mathcal{O}'$. The electric field $\mathbf{E}^\epsilon \equiv \mathbf{E}^\epsilon(t,\mathbf{x})$ standing in the Vlasov equation is now given by the Poisson equation, and then the system we work with writes

$$
\begin{aligned}
&\frac{\partial f^\epsilon}{\partial t} + \frac{1}{\epsilon}\mathbf{v}\cdot\nabla_x f^\epsilon + \left(\mathbf{E}^\epsilon + \frac{1}{\epsilon}\mathbf{v}\times\mathbf{m}\right)\cdot\nabla_v f^\epsilon = 0,\\
&f^\epsilon_{|t=0} = f_0,\\
&\mathbf{E}^\epsilon = -\nabla\phi^\epsilon, \quad -\Delta\phi^\epsilon = \rho^\epsilon,\\
&\rho^\epsilon = \int_{\mathbb{R}^2_v} f^\epsilon \, d\mathbf{v}.
\end{aligned}
\tag{1.17}
$$

As the first equation in (1.17) is bidimensional, we precise the sense to give to $\mathbf{v}\times\mathbf{m}$

$$
\mathbf{v}\times\mathbf{m} = \begin{pmatrix} v_3 \\ -v_2 \end{pmatrix}. \tag{1.18}
$$

We assume

$$
f_0 \geq 0, \quad f_0 \in L^1 \cap L^p(\mathcal{O}), \quad 0 < \int_{\mathcal{O}} f_0(1+|\mathbf{v}|^2)\, d\mathbf{v} < +\infty, \tag{1.19}
$$

for some $p \geq 2$, and we have the following theorem.

THEOREM 1.4. *Under assumption* (1.19)*, for each* ϵ*, there exists a solution* $(f^\epsilon, \mathbf{E}^\epsilon)$ *of* (1.17) *in* $L^\infty(0,T; L^1 \cap L^p(\mathcal{O})) \times L^\infty(0,T; W^{1,\frac{3}{2}}(\mathbb{R}^2_x))$ *for any* $T \in \mathbb{R}^+$*. Moreover, this solution is bounded in* $L^\infty(0,T; L^1 \cap L^p(\mathcal{O})) \times L^\infty(0,T; W^{1,\frac{3}{2}}(\mathbb{R}^2_x))$ *independently on* ϵ*.*

If we consider a sequence $(f^\epsilon, \mathbf{E}^\epsilon)$ *of such solutions, extracting a subsequence, we have*

$$
\begin{aligned}
&f^\epsilon \text{ 2-scale converges to } F \in L^\infty(0,T; L^\infty_{2\pi}(\mathbb{R}_\tau; L^p(\mathcal{O}))),\\
&\mathbf{E}^\epsilon \text{ 2-scale converges to } \mathcal{E} \in L^\infty(0,T; L^\infty_{2\pi}(\mathbb{R}_\tau; W^{1,\frac{3}{2}}(\mathbb{R}^2_x))),
\end{aligned}
\tag{1.20}
$$

where $F \equiv F(t,\tau,\mathbf{x},\mathbf{v})$ *and* $\mathcal{E} \equiv \mathcal{E}(t,\tau,\mathbf{x})$*.*

Moreover, there exists a function $G \equiv G(t,\mathbf{y},\mathbf{u}) \in L^\infty(0,T; L^1 \cap L^p(\mathcal{O}'))$ *such that*

$$
F(t,\tau,\mathbf{x},\mathbf{v}) = G(t, \mathbf{x} + \mathcal{R}(-\tau)\mathbf{v}, R(-\tau)\mathbf{v}), \tag{1.21}
$$

and $(G,\mathcal{E})$ *is solution of*

(1.22)
$$\frac{\partial G}{\partial t}+\frac{1}{2\pi}\left(\int_0^{2\pi}\mathcal{R}(-\tau)\mathcal{E}(t,\tau,\mathbf{y}+\mathcal{R}(\tau)\mathbf{u})\,d\tau\right)\cdot\nabla_y G + \frac{1}{2\pi}\left(\int_0^{2\pi}R(-\tau)\mathcal{E}(t,\tau,\mathbf{y}+\mathcal{R}(\tau)\mathbf{u})\,d\tau\right)\cdot\nabla_u G=0,$$

$$G_{|t=0}=\frac{1}{2\pi}f_0,$$

$$\mathcal{E}\equiv\mathcal{E}(t,\tau,\mathbf{x}),\ \textit{with}\ \mathcal{E}=-\nabla\Phi,\ \ -\Delta\Phi=\int G(t,\mathbf{x}+\mathcal{R}(-\tau)\mathbf{v},R(-\tau)\mathbf{v})\,d\mathbf{v},$$

with R *and* $\mathcal{R}$ *given by*

$$R(\tau)=\begin{pmatrix}\cos\tau & \sin\tau\\ -\sin\tau & \cos\tau\end{pmatrix},\ \ \mathcal{R}(\tau)=\begin{pmatrix}\sin\tau & 1-\cos\tau\\ \cos\tau-1 & \sin\tau\end{pmatrix}. \tag{1.23}$$

In order to prove this theorem, we modify the generic framework previously introduced. We consider here

$$\begin{aligned}&\frac{\partial u^\epsilon}{\partial t}+\mathbf{A}^\epsilon\cdot\nabla_x u^\epsilon+\frac{1}{\epsilon}\mathbf{L}\cdot\nabla_x u^\epsilon=0,\\ &u^\epsilon_{t=0}=u_0,\end{aligned} \tag{1.24}$$

where the notations are similar as for (1.10): $u^\epsilon\equiv u^\epsilon(t,\mathbf{x})$, $t\in[0,T)$, $T<\infty$; $\mathbf{x}\in\mathbb{R}^n=\mathcal{O}$, $\mathcal{Q}=[0,T)\times\mathcal{O}$. We suppose, as previously, that $\mathbf{L}\equiv M\mathbf{x}+N$, where M is a constant entry matrix satisfying $\mathrm{tr}M=0$ and $e^{\tau M}$ is θ-periodic and where $N\in\mathrm{Im}M$. The assumptions we make on $\mathbf{A}^\epsilon$ are the following: we suppose that, for all $\epsilon>0$, $\nabla_x\cdot\mathbf{A}^\epsilon=0$ and that, for some $q>1$,

$$\mathbf{A}^\epsilon\ \text{2-scale converges to}\ \mathcal{A}\in L^\infty(0,T;L^\infty_\theta(\mathbb{R}_\tau;W^{1,q}(K))) \tag{1.25}$$

for all compact sets $K\subset\mathbb{R}^n$ and where $\mathcal{A}\equiv\mathcal{A}(t,\tau,\mathbf{x})$ is θ-periodic in τ.

We have the following.

THEOREM 1.5. *Under the assumptions above, if, moreover, the sequence* (u^ϵ) *of solutions of* (1.24) *satisfies*

$$\|u^\epsilon\|_{L^\infty(0,T;L^p(\mathcal{O}))}\le C, \tag{1.26}$$

for some $p>1$ *such that* $\frac{1}{p}+\frac{1}{q'}<1$, *where* $\frac{1}{q'}=\mathrm{Max}\{\frac{1}{q}-\frac{1}{n},0\}$; *then, extracting a subsequence,*

$$u^\epsilon\ \textit{2-scale converges to a profile}\ U\in L^\infty(0,T;L^\infty_\theta(\mathbb{R}_\tau;L^p(\mathcal{O}))). \tag{1.27}$$

Moreover, we have

$$U(t,\tau,\mathbf{x})=U_0(t,e^{-\tau M}(\mathbf{x}-\overline{N})+\overline{N}), \tag{1.28}$$

where $\overline{N}$ *is such that* $-M\overline{N}=N$ *and where* $U_0\equiv U_0(t,\mathbf{y})$ *is solution of*

$$\begin{aligned}&\frac{\partial U_0}{\partial t}+\int_0^\theta e^{-\sigma M}\mathcal{A}(t,\sigma,e^{\sigma M}(\mathbf{y}-\overline{N})+\overline{N})\,d\sigma\cdot\nabla_y U_0=0,\\ &U_{0|t=0}=\tfrac{1}{\theta}u_0.\end{aligned} \tag{1.29}$$

The proof of this theorem consists in finding the constraint equation imposed on U by the $\frac{1}{\epsilon}L$ operator. This yields (1.28). Then we remove the essential oscillation of u^ϵ by defining $w^\epsilon(t,\mathbf{y}) = u^\epsilon(t, e^{\frac{t}{\epsilon}M}(\mathbf{y}-\overline{N})+\overline{N})$. Using the equation w^ϵ satisfies, denoting $(W_0^{1,r}(K))^*$ the dual of $W_0^{1,r}(K)$, we prove that $\frac{\partial w^\epsilon}{\partial t}$ is bounded in $L^\infty(0,T;(W_0^{1,r}(K))^*)$, for some $r>1$ ($\frac{1}{r^*} = \frac{1}{p}+\frac{1}{q}-\frac{1}{n}$, $\frac{1}{r}+\frac{1}{r^*}=1$), which, applying the Aubin–Lions lemma, gives that $w^\epsilon \to \theta U_0$ strongly in $L^\infty(0,T;(W_0^{1,q}(K))^*)$ for any compact set $K\subset\mathbb{R}^n$. This fact, coupled with (1.25), enables us to pass to the limit in the equation satisfied by w^ϵ and find (1.29).

Theorem 1.4 is a direct application of Theorem 1.5 once the wanted regularity of $\mathbf{E}^\epsilon$ is proved. This is done with the help of classical kinetic energy estimates and the regularization property of the Laplace operator.

The paper is organized as follows. In section 2 we present the scaling leading to the finite Larmor radius approximation. We show how to obtain (1.1) and system (1.17). The next section is devoted to the deduction of the asymptotic behavior of the linear Vlasov equation. Finally, in section 4 we prove Theorems 1.5 and 1.4 concerning the nonlinear case.

2. Scaling: The finite Larmor radius regime. Approximate models in the case of a large external magnetic field have been used by physicists for a long time and the corresponding gyrokinetic ordering is due to Taylor and Hastie [24] and Rutherford and Frieman [19]. We also refer to [6] for a further discussion. And for a physical introduction of the finite Larmor radius model, we refer to [15, 17]. Our scaling assumptions follow from those works.

We present here the scaling leading to (1.1) and system (1.17). We exhibit the important parameters playing a role when charged particles are submitted to a strong magnetic field. For this purpose we consider the following Vlasov–Poisson system

$$
\begin{aligned}
&\frac{\partial f}{\partial t} + \mathbf{v}\cdot\nabla_x f + \frac{q}{m}(\mathbf{E}(t,\mathbf{x}) + \mathbf{v}\times\mathbf{B}(t,\mathbf{x}))\cdot\nabla_v f = 0\\
&f_{|t=0} = f_0,\\
&\mathbf{E} = -\nabla\phi, \qquad -\Delta\phi = \frac{q}{\epsilon_0}\rho,\\
&\rho(t,\mathbf{x}) = \int_{\mathbb{R}^3_v} f(t,\mathbf{x},\mathbf{v})\,d\mathbf{v},
\end{aligned}
\tag{2.1}
$$

before any scaling, which can be considered as a natural model to describe the behavior of charged particles under the action of an external magnetic field $\mathbf{B}(t,\mathbf{x})$.

We define some characteristic scales: $\bar{t}$ stands for a characteristic time, $\overline{L_\parallel}$ for a characteristic length in the direction of the magnetic field, $\overline{L_\perp}$ for a characteristic length in the direction orthogonal to the magnetic field, $\overline{v}$ for a characteristic velocity. Denoting, for any vector $\mathbf{x}$, $\mathbf{x}_\parallel$ and $\mathbf{x}_\perp$ its components parallel and perpendicular to the magnetic field, we now define new variables t', $\mathbf{x}'$, and $\mathbf{v}'$, by $t = \bar{t}t'$, $\mathbf{x}_\parallel = \overline{L_\parallel}\mathbf{x}'_\parallel$, $\mathbf{x}_\perp = \overline{L_\perp}\mathbf{x}'_\perp$, and $\mathbf{v} = \overline{v}\mathbf{v}'$, making the characteristic scales the unities. In the same way, we define the scaling factors for the fields: $\overline{E}$ for the electric field and $\overline{B}$ for the magnetic field and the new fields $\mathcal{E}$ and $\mathcal{B}$ are given by $\overline{E}\mathcal{E}(t',\mathbf{x}') = \mathbf{E}(\bar{t}t', \overline{L_\parallel}\mathbf{x}'_\parallel, \overline{L_\perp}\mathbf{x}'_\perp)$ and $\overline{B}\mathcal{B}(t',\mathbf{x}') = \mathbf{B}(\bar{t}t', \overline{L_\parallel}\mathbf{x}'_\parallel, \overline{L_\perp}\mathbf{x}'_\perp)$. Lastly, defining a scaling factor $\overline{f}$ for the repartition function, noticing that f is a repartition function on the phase-space, it is natural to

define the new repartition function by

$$\overline{f} f'(t', \mathbf{x}', \mathbf{v}') = \overline{L_\parallel}\,\overline{L_\perp}^2 \overline{v}^3 f(\overline{t}t', \overline{L_\parallel}\mathbf{x}'_\parallel, \overline{L_\perp}\mathbf{x}'_\perp, \overline{v}\mathbf{v}'). \tag{2.2}$$

With those new variables and fields we deduce the scaling equations.

2.1. Scaling of the Vlasov equation. Let us begin with the Vlasov equation; we obtain that f' is solution of

$$\frac{\partial f'}{\partial t'} + \frac{\overline{v}\overline{t}}{\overline{L_\parallel}}\mathbf{v}'_\parallel \cdot \nabla_{x'} f' + \frac{\overline{v}\overline{t}}{\overline{L_\perp}}\mathbf{v}'_\perp \cdot \nabla_{x'} f' + \left(\frac{q\overline{E}\overline{t}}{m\overline{v}}\mathcal{E}(t', \mathbf{x}') + \frac{q\overline{B}\overline{t}}{m}\mathbf{v}' \times \mathcal{B}(t', \mathbf{x}')\right) \cdot \nabla_{v'} f' = 0. \tag{2.3}$$

Now, we introduce the characteristic cyclotron frequency: $\overline{\omega}_c = \frac{q\overline{B}}{m}$ and the characteristic Larmor radius: $\overline{a}_L = \frac{\overline{v}}{\overline{\omega}_c}$. Using those physical quantities, (2.3) becomes

$$\frac{\partial f'}{\partial t'} + \overline{t\omega}_c \frac{\overline{a}_L}{\overline{L_\parallel}}\mathbf{v}'_\parallel \cdot \nabla_{x'} f' + \overline{t\omega}_c \frac{\overline{a}_L}{\overline{L_\perp}}\mathbf{v}'_\perp \cdot \nabla_{x'} f' + \left(\overline{t\omega}_c \frac{\overline{E}}{\overline{v}\overline{B}}\mathcal{E}(t', \mathbf{x}') + \overline{t\omega}_c \mathbf{v}' \times \mathcal{B}(t', \mathbf{x}')\right) \cdot \nabla_{v'} f' = 0. \tag{2.4}$$

Assuming the magnetic field is strong consists essentially in setting

$$\overline{t\omega_c} = \frac{1}{\epsilon} \text{ and } \frac{\overline{E}}{\overline{v}\overline{B}} = \epsilon \tag{2.5}$$

for a small parameter ϵ, and the finite Larmor radius regime consists in choosing

$$\frac{\overline{a}_L}{\overline{L_\parallel}} = \epsilon \text{ and } \frac{\overline{a}_L}{\overline{L_\perp}} = 1. \tag{2.6}$$

Hence the rescaled Vlasov equation writes

$$\frac{\partial f'}{\partial t'} + \mathbf{v}'_\parallel \cdot \nabla_{x'} f' + \frac{1}{\epsilon}\mathbf{v}'_\perp \cdot \nabla_{x'} f' + (\mathcal{E}(t', \mathbf{x}') + \frac{1}{\epsilon}\mathbf{v} \times \mathcal{B}(t', \mathbf{x}')) \cdot \nabla_{v'} f' = 0. \tag{2.7}$$

Concerning the initial data, under the scaling (2.2), the second equation of (2.1) directly gives

$$f'_{|t'=0} = \frac{\overline{L_\parallel}\,\overline{L_\perp}^2 \overline{v}^3}{\overline{f}} f_0(\overline{L_\parallel}\mathbf{x}'_\parallel, \overline{L_\perp}\mathbf{x}'_\perp, \overline{v}\mathbf{v}'). \tag{2.8}$$

Hence, if we assume that the scales of variations of the initial data f_0 (before scaling) are of the same order as the characteristic lengths used, and that $\overline{f} = \overline{L_\parallel}\,\overline{L_\perp}^2 \overline{v}^3$, it is natural to consider (1.1) as a relevant model to understand local behavior of charged particles under the action of a strong external constant magnetic field.

This is the reason why we study (1.1) in section 3.

2.2. Scaling of the Poisson equation. We now turn to the Poisson equation given by the third and fourth equations of (2.1). For this purpose, we define the new electric potential by

$$\overline{E}\,\overline{L_\parallel}\phi'(t', \mathbf{x}') = \phi(\overline{t}t', \overline{L_\parallel}\mathbf{x}'_\parallel, \overline{L_\perp}\mathbf{x}'_\perp), \tag{2.9}$$

and the new particle density by

$$\rho'(t',\mathbf{x}') = \int f'(t',\mathbf{x}',\mathbf{v}')\,d\mathbf{v}'. \tag{2.10}$$

Direct computations give

$$\rho'(t',\mathbf{x}') = \frac{\overline{L_\parallel}\,\overline{L_\perp}^2}{\overline{f}}\rho(\bar{t}t',\overline{L_\parallel}\mathbf{x}'_\parallel,\overline{L_\perp}\mathbf{x}'_\perp), \tag{2.11}$$

$$\nabla\phi(\bar{t}t',\overline{L_\parallel}\mathbf{x}'_\parallel,\overline{L_\perp}\mathbf{x}'_\perp) = \overline{E}\begin{pmatrix}\nabla_{\mathbf{x}'_\parallel}\phi'(t',\mathbf{x}')\\ \frac{\overline{L_\parallel}}{\overline{L_\perp}}\nabla_{\mathbf{x}'_\perp}\phi'(t',\mathbf{x}')\end{pmatrix}, \tag{2.12}$$

and

$$\Delta\phi(\bar{t}t',\overline{L_\parallel}\mathbf{x}'_\parallel,\overline{L_\perp}\mathbf{x}'_\perp) = \frac{\overline{E}}{\overline{L_\parallel}}\left(\Delta_{\mathbf{x}'_\parallel}\phi'(t',\mathbf{x}') + \frac{\overline{L_\parallel}^2}{\overline{L_\perp}^2}\Delta_{\mathbf{x}'_\perp}\phi'(t',\mathbf{x}')\right). \tag{2.13}$$

Hence the Poisson equation $-\Delta\phi = \frac{q}{\epsilon_0}\rho$ becomes

$$-\left(\Delta_{\mathbf{x}'_\parallel}\phi' + \frac{\overline{L_\parallel}^2}{\overline{L_\perp}^2}\Delta_{\mathbf{x}'_\perp}\phi'\right) = \frac{q}{\epsilon_0}\frac{\overline{f}}{\overline{E}\,\overline{L_\perp}^2}\rho' \tag{2.14}$$

and the definition of the electric field $\mathbf{E} = -\nabla\phi$ yields

$$\mathcal{E} = -\begin{pmatrix}\nabla_{\mathbf{x}'_\parallel}\phi'\\ \frac{\overline{L_\parallel}}{\overline{L_\perp}}\nabla_{\mathbf{x}'_\perp}\phi'\end{pmatrix}. \tag{2.15}$$

Setting now the same ratio as in (2.5) and (2.6) and considering that the scales of variations of the initial data are of the same order as the characteristic lengths, the rescaled Vlasov–Poisson system writes

$$\begin{aligned}
&\frac{\partial f'}{\partial t'} + \mathbf{v}'_\parallel\cdot\nabla_{x'}f' + \frac{1}{\epsilon}\mathbf{v}'_\perp\cdot\nabla_{x'}f' + \left(\mathcal{E}(t',\mathbf{x}') + \frac{1}{\epsilon}\mathbf{v}\times\mathcal{B}(t',\mathbf{x}')\right)\cdot\nabla_{v'}f' = 0,\\
&f_{|t=0} = f'_0,\\
&\mathcal{E} = -\begin{pmatrix}\nabla_{\mathbf{x}'_\parallel}\phi'\\ \frac{1}{\epsilon}\nabla_{\mathbf{x}'_\perp}\phi'\end{pmatrix},\quad -\left(\Delta_{\mathbf{x}'_\parallel}\phi' + \frac{1}{\epsilon^2}\Delta_{\mathbf{x}'_\perp}\phi'\right) = \gamma\rho',\\
&\rho'(t',\mathbf{x}') = \int_{\mathbb{R}^3_v} f'(t',\mathbf{x}',\mathbf{v}')\,d\mathbf{v}',
\end{aligned} \tag{2.16}$$

with $\gamma = \frac{q}{\epsilon_0}\frac{\overline{f}}{\overline{E}\overline{L_\perp}^2}$.

For the study we lead in section 4 we consider the previous system with $\gamma = \frac{1}{\epsilon}$ and with $\mathcal{B} = \mathbf{m} = \mathbf{e}_1$. We moreover assume that none of the fields depend on the component parallel to the magnetic fields $\mathbf{x}_\parallel$ and $\mathbf{v}_\parallel$. In this case the Poisson equation from which we remove the $\mathbf{x}_\parallel-$dependency

$$\mathcal{E} = -\begin{pmatrix}0\\ \frac{1}{\epsilon}\nabla_{\mathbf{x}'_\perp}\phi'\end{pmatrix},\quad -\frac{1}{\epsilon^2}\Delta_{\mathbf{x}'_\perp}\phi' = \frac{1}{\epsilon}\rho', \tag{2.17}$$

is equivalent to, removing the magnetic field direction,

$$\mathcal{E} = -\nabla\phi^*, \quad -\Delta\phi^* = \rho^*, \tag{2.18}$$

where ϕ^* is nothing but $\frac{1}{\epsilon}\phi'$ and with

$$\rho^* = \int_{\mathbb{R}^2_v} f'\, d\mathbf{v}, \tag{2.19}$$

explaining the interest of studying system (1.17).

3. Homogenization of the Vlasov equation. In this section, we provide the homogenization of the Vlasov equation (1.1) and prove Theorem 1.1. Since the contexts of (1.1) and the equation studied in Frénod and Sonnendrücker [8] are similar, we develop a generic framework and apply it to prove Theorem 1.1. We then show that this generic framework applies also to prove Theorems 1.1 and 3.2 of Frénod and Sonnendrücker [8].

3.1. Generic framework—proof of Theorem 1.3. The framework inside which the problem we want to homogenize enters is the following conservation law singularly linearly perturbed:

$$\begin{aligned} &\frac{\partial u^\epsilon}{\partial t} + \mathbf{A}\cdot\nabla_x u^\epsilon + \frac{1}{\epsilon}\mathbf{L}\cdot\nabla_x u^\epsilon = 0,\\ &u^\epsilon_{|t=0} = u_0, \end{aligned} \tag{3.1}$$

where $u^\epsilon \equiv u^\epsilon(t,\mathbf{x})$, $t \in [0,T)$ for some $T < \infty$, and $\mathbf{x} \in \mathbb{R}^n = \mathcal{O}$. We denote $\mathcal{Q} = [0,T)\times\mathcal{O}$, and we assume $\mathbf{A} \equiv \mathbf{A}(t,\mathbf{x}) \in L^\infty(0,T;L^2_{loc}(\mathcal{O}))$, with $\nabla_x\cdot\mathbf{A} = 0$ and $\mathbf{L} \equiv M\mathbf{x} + N$, where M is a real $n\times n$ matrix with constant entries satisfying $\mathrm{tr}M = 0$ and where $N \in \mathrm{Im}M$, which implies that $\nabla_x\cdot\mathbf{L} = 0$.We moreover assume that $e^{\tau M}$ is θ-periodic for a given $\theta\in\mathbb{R}$.

The proof of Theorem 1.3, characterizing the limit of (3.1), is led in three steps. First, we look for the constraint imposed by the operator $(\frac{1}{\epsilon}\mathbf{L}\cdot\nabla_x)$ on the profile U, 2-scale limit of (u^ϵ). Studying the characteristics associated with this constraint, we obtain the form (1.14) it gives to U. In the second step, using test functions satisfying the constraint in the weak formulation of (3.1), we get the equation satisfied by U_0. In view of formula (1.8) linking the 2-scale limit to the weak$-*$ limit, in the last step, we integrate the equation satisfied by U_0 to deduce (1.16).

Under the assumption (1.11), we may apply the result of Nguetseng [18] and Allaire [2] (see Theorem 1.2). Then, for any period $\tilde\theta$ there exists a $\tilde\theta$-periodic profile $U_{\tilde\theta}(t,\tau,\mathbf{x}) \in L^\infty(0,T;L^\infty_{\tilde\theta}(\mathbb{R}_\tau;L^2(\mathcal{O})))$ such that, for any regular function $\psi_{\tilde\theta}(t,\tau,\mathbf{x})$ compactly supported in $(t,\mathbf{x})$ and $\tilde\theta$-periodic in τ, we have

$$\int_{\mathcal{Q}} u^\epsilon(t,\mathbf{x})\psi_{\tilde\theta}\left(t,\frac{t}{\epsilon},\mathbf{x}\right)\,dt\,d\mathbf{x} \to \int_{\mathcal{Q}}\int_0^{\tilde\theta} U(t,\tau,\mathbf{x})\psi_{\tilde\theta}(t,\tau,\mathbf{x})\,d\tau\,dt\,d\mathbf{x}. \tag{3.2}$$

Now, we write a weak formulation of (3.1) with oscillating test functions $(\psi_{\tilde\theta})^\epsilon = \psi_{\tilde\theta}(t,\frac{t}{\epsilon},\mathbf{x})$, with $\psi_{\tilde\theta}(t,\tau,\mathbf{x})$ previously defined. Since $\nabla_x\cdot\mathbf{A} = \nabla_x\cdot\mathbf{L} = 0$, it writes

$$\int_{\mathcal{Q}} u^\epsilon\left(\left(\frac{\partial\psi_{\tilde\theta}}{\partial t}\right)^\epsilon + \frac{1}{\epsilon}\left(\frac{\partial\psi_{\tilde\theta}}{\partial\tau}\right)^\epsilon + \mathbf{A}\cdot(\nabla_x\psi_{\tilde\theta})^\epsilon + \frac{1}{\epsilon}\mathbf{L}\cdot(\nabla_x\psi_{\tilde\theta})^\epsilon\right)dtd\mathbf{x} = -\int_{\mathcal{O}} u_0\psi_{\tilde\theta}(0,0,\mathbf{x})\,d\mathbf{x}. \tag{3.3}$$

Multiplying (3.3) by ϵ and passing to the limit gives the following constraint equation for the $\tilde{\theta}$-periodic profile $U_{\tilde{\theta}}$:

$$\frac{\partial U_{\tilde{\theta}}}{\partial \tau} + \mathbf{L} \cdot \nabla_x U_{\tilde{\theta}} = 0 \text{ in } \mathcal{D}'(\mathbb{R}_\tau \times \mathcal{O}). \tag{3.4}$$

This equation says that $U_{\tilde{\theta}}$ is constant along the characteristics of the following dynamical system:

$$\frac{d\mathbf{X}}{d\tau} = \mathbf{L}(\mathbf{X}(\tau)) = M\mathbf{X}(\tau) + N. \tag{3.5}$$

Using the assumptions made on $\mathbf{L}$, writing $\mathbf{X}(\tau; \mathbf{x}, s)$ for the solution of (3.5) satisfying $\mathbf{X}(s; \mathbf{x}, s) = \mathbf{x}$, we obtain

$$\mathbf{X}(\tau; \mathbf{x}, s) = e^{(\tau - s)M}(\mathbf{x} - \overline{N}) + \overline{N}. \tag{3.6}$$

Hence from (3.4), we deduce, on the one hand, that for any $\tilde{\theta}$ the $\tilde{\theta}$-periodic profile writes

$$U_{\tilde{\theta}}(t, \tau, \mathbf{x}) = U_0(t, e^{-\tau M}(\mathbf{x} - \overline{N}) + \overline{N}) \tag{3.7}$$

for a function $U_0 \equiv U_0(t, \mathbf{y}) \in L^\infty(0, T; L^2(\mathcal{O}'))$. On the other hand, we take the θ-periodicity of $e^{-\tau M}$ under consideration. In view of (3.7), we deduce that if $\tilde{\theta}$ and θ are incommensurable, $U_{\tilde{\theta}}$ cannot depend on τ, and then contains no information concerning the oscillations of (u^ϵ). Yet if $\tilde{\theta}$ equals (or is multiple of) θ, the profile $U_{\tilde{\theta}}$ naturally satisfies its $\tilde{\theta}$-periodicity condition once (3.7) is satisfied.

Hence, among every possible periodic profile, we are incited to work now with the θ-periodic one

$$U := U_\theta, \tag{3.8}$$

which writes

$$U(t, \tau, \mathbf{x}) = U_0(t, e^{-\tau M}(\mathbf{x} - \overline{N}) + \overline{N}) \tag{3.9}$$

for $U_0 \equiv U_0(t, \mathbf{y}) \in L^\infty(0, T; L^2(\mathcal{O}'))$, which is the equality (1.14) of Theorem 1.3.

Now, we seek the equation U_0 satisfies. For this purpose, we build oscillating test functions satisfying the constraint and use them in the weak formulation (3.3).

For any $\varphi(t, \mathbf{y})$, regular and compactly supported, we define $\psi(t, \tau, \mathbf{x}) = \varphi(t, e^{-\tau M}(\mathbf{x} - \overline{N}) + \overline{N})$ and we inject in (3.3) test function $(\psi)^\epsilon = \psi(t, \frac{t}{\epsilon}, \mathbf{x})$. Acting in such a way, the terms containing the constraint vanishes. We have then

$$\int_{\mathcal{Q}} u^\epsilon \left(\left(\frac{\partial \psi}{\partial t} \right)^\epsilon + \mathbf{A} \cdot (\nabla_x \psi)^\epsilon \right) dt d\mathbf{x} = - \int_{\mathcal{O}} u_0 \psi(0, 0, \mathbf{x})\, d\mathbf{x}, \tag{3.10}$$

which passing to the limit yields

$$\int_{\mathcal{Q}} \int_0^\theta U \left(\frac{\partial \psi}{\partial t} + \mathbf{A} \cdot \nabla \psi \right) dt d\mathbf{x} d\tau = - \int_{\mathcal{O}} u_0 \psi(0, 0, \mathbf{x})\, d\mathbf{x}. \tag{3.11}$$

In (3.11) we use the expression of U in terms of U_0, the expression of ψ in term of φ, without forgetting

$$\nabla_x \psi(t, \tau, \mathbf{x}) = (e^{-\tau M})^T \nabla_y \varphi(t, e^{-\tau M}(\mathbf{x} - \overline{N}) + \overline{N}), \tag{3.12}$$

denoting $(e^{-\tau M})^T$ the transpose of $e^{-\tau M}$; and we make the change of variable $\mathbf{x} \mapsto \mathbf{y} = e^{-\tau M}(\mathbf{x} - \overline{N}) + \overline{N}$. This gives

$$\int_{\mathcal{Q}'} \int_0^\theta U_0 \left(\frac{\partial \varphi}{\partial t} + e^{-\tau M} \mathbf{A}(t, e^{\tau M}(\mathbf{y} - \overline{N}) + \overline{N}) \cdot \nabla_y \varphi \right) dt d\mathbf{y} d\tau = - \int_{\mathcal{O}} u_0 \varphi(0, \mathbf{y})\, d\mathbf{y}. \tag{3.13}$$

An easy computation coupled with the fact that $\nabla_x \cdot \mathbf{A} = 0$ gives

$$\nabla_y \cdot (e^{-\sigma M} \mathbf{A}(t, e^{\sigma M}(\mathbf{y} - \overline{N}) + \overline{N})) = (\nabla_x \cdot \mathbf{A})(t, e^{\sigma M}(\mathbf{y} - \overline{N}) + \overline{N}) = 0. \tag{3.14}$$

Hence, knowing that neither U_0 nor φ depend on τ, we deduce that (3.13) is the weak formulation of

$$\begin{aligned} &\frac{\partial U_0}{\partial t} + \frac{1}{\theta} \int_0^\theta e^{-\sigma M} \mathbf{A}(t, e^{\sigma M}(\mathbf{y} - \overline{N}) + \overline{N})\, d\sigma \cdot \nabla_y U_0 = 0, \\ &U_{0|t=0} = \frac{1}{\theta} u_0, \end{aligned} \tag{3.15}$$

proving the second part of Theorem 1.3.

In order to get (1.16) we use the fact that

$$u(t, \mathbf{x}) = \int_0^\theta U(t, \tau, \mathbf{x})\, d\tau = \int_0^\theta U_0(t, e^{-\tau M}(\mathbf{x} - \overline{N}) + \overline{N})\, d\tau. \tag{3.16}$$

Replacing in (3.15) $\mathbf{y}$ by $e^{-\tau M}(\mathbf{x} - \overline{N}) + \overline{N}$ and integrating in τ we get

$$\begin{aligned} &\frac{\partial}{\partial t} \left(\int_0^\theta U_0(t, e^{-\tau M}(\mathbf{x} - \overline{N}) + \overline{N})\, d\tau \right) \\ &\quad + \frac{1}{\theta} \int_0^\theta \int_0^\theta e^{-\sigma M} \mathbf{A}(t, e^{(\sigma - \tau) M}(\mathbf{x} - \overline{N}) + \overline{N})\, d\sigma \cdot \nabla_y U_0(t, e^{-\tau M}(\mathbf{x} - \overline{N}) + \overline{N})\, d\tau = 0, \\ &\int_0^\theta U_0(t, e^{-\tau M}(\mathbf{x} - \overline{N}) + \overline{N}))_{|t=0}\, d\tau = \frac{1}{\theta} \int_0^\theta u_0(e^{-\tau M}(\mathbf{x} - \overline{N}) + \overline{N})\, d\tau. \end{aligned} \tag{3.17}$$

An easy computation yields

$$\nabla_x (U_0(t, e^{-\tau M}(\mathbf{x} - \overline{N}) + \overline{N})) = (e^{-\tau M})^T (\nabla_y U_0)(t, e^{-\tau M}(\mathbf{x} - \overline{N}) + \overline{N}), \tag{3.18}$$

and then replacing in the second term of the first equation in (3.17) gives

$$\begin{aligned} &\int_0^\theta \int_0^\theta e^{-\sigma M} \mathbf{A}(t, e^{(\sigma - \tau) M}(\mathbf{x} - \overline{N}) + \overline{N})\, d\sigma \cdot (e^{\tau M})^T \nabla_x (U_0(t, e^{-\tau M}(\mathbf{x} - \overline{N}) + \overline{N})\, d\tau \\ &\quad = \int_0^\theta \int_0^\theta e^{(\tau - \sigma) M} \mathbf{A}(t, e^{(\sigma - \tau) M}(\mathbf{x} - \overline{N}) + \overline{N})\, d\sigma \cdot \nabla_x (U_0(t, e^{-\tau M}(\mathbf{x} - \overline{N}) + \overline{N}))\, d\tau. \end{aligned} \tag{3.19}$$

Yet by the periodicity of $\tau \mapsto e^{\tau M}$ we deduce that

$$\int_0^\theta e^{(\tau - \sigma) M} \mathbf{A}(t, e^{(\sigma - \tau) M}(\mathbf{x} - \overline{N}) + \overline{N})\, d\sigma = \int_0^\theta e^{-\sigma M} \mathbf{A}(t, e^{\sigma M}(\mathbf{x} - \overline{N}) + \overline{N})\, d\sigma \tag{3.20}$$

does not depend on τ.

We may finally conclude that (3.17) reads

$$\begin{aligned} &\frac{\partial u}{\partial t} + \frac{1}{\theta}\int_0^\theta e^{-\sigma M}\mathbf{A}(t, e^{\sigma M}(\mathbf{x}-\overline{N})+\overline{N})d\sigma \cdot \nabla u = 0, \\ &u_{|t=0}(\mathbf{x}) = \frac{1}{\theta}\int_0^\theta u_0(e^{-\sigma M}(\mathbf{x}-\overline{N})+\overline{N})\,d\sigma. \end{aligned} \tag{3.21}$$

achieving the proof of Theorem 1.3.

3.2. Application to the Vlasov equation—proof of Theorem 1.1. Using assumption (1.2) made on f_0 and the following property of f^ϵ solution of (1.1)

$$\frac{d}{dt}\|f^\epsilon(t,\cdot,\cdot)\|_{L^2(\mathcal{O})} = 0, \tag{3.22}$$

obtained by a direct integration in $\mathbf{x}$ and $\mathbf{v}$ of the first equation in (1.1), after multiplication by f^ϵ, we deduce that

$$\|f^\epsilon\|_{L^\infty(0,T;L^2(\mathcal{O}))} \leq C \tag{3.23}$$

for some constants C.

Hence, the Vlasov equation (1.1) enters the generic framework previously built with

$$\mathbf{A}(t,\mathbf{x},\mathbf{v}) = \begin{pmatrix} \mathbf{v}_\parallel \\ \mathbf{E}(t,\mathbf{x}) \end{pmatrix} (\in \mathbb{R}^6) \text{ and } \mathbf{L}(t,\mathbf{x},\mathbf{v}) = \begin{pmatrix} \mathbf{v}_\perp \\ \mathbf{v}\times\mathbf{m} \end{pmatrix} (\in \mathbb{R}^6). \tag{3.24}$$

Then the differential system defining the characteristics $(\dot{X},\dot{V}) = \mathbf{L}(X,V)$ becomes

$$\frac{dX_\perp}{d\tau} = V_\perp, \qquad \frac{dV}{d\tau} = V\times\mathbf{m}.$$

An easy computation then yields $V(\tau;\mathbf{v},s) = R(\tau-s)\mathbf{v}$ and $X(\tau;(\mathbf{x},\mathbf{v}),s) = \mathbf{x} + \mathcal{R}(\tau-s)\mathbf{v}$, with $\mathcal{R}(\tau)$ and $R(\tau)$ given by (1.6). Hence $e^{\tau M}$ reads in this case

$$e^{\tau M} = \begin{pmatrix} I & \mathcal{R}(\tau) \\ 0 & R(\tau) \end{pmatrix}. \tag{3.25}$$

We can then deduce

$$f^\epsilon \text{ 2-scale converges to } F \in L^\infty(0,T;L^\infty_{2\pi}(\mathbb{R}_\tau;L^2(\mathcal{O}))), \tag{3.26}$$

and applying Theorem 1.3, we can deduce that there exists a function $G \equiv G(t,\mathbf{y},\mathbf{u}) \in L^\infty(0,T;L^2(\mathcal{O}'))$ such that

$$F(t,\tau,\mathbf{x},\mathbf{v}) = G(t,\mathbf{x}+\mathcal{R}(-\tau)\mathbf{v}, R(-\tau)\mathbf{v}), \tag{3.27}$$

where G is the unique solution of

$$\begin{aligned} &\frac{\partial G}{\partial t} + \mathbf{u}_\parallel\cdot\nabla_y G + \frac{1}{2\pi}\left(\int_0^{2\pi}\mathcal{R}(-\tau)\mathbf{E}(t,\mathbf{y}+\mathcal{R}(\tau)\mathbf{u})\,d\tau\right)\cdot\nabla_y G \\ &\qquad + \frac{1}{2\pi}\left(\int_0^{2\pi} R(-\tau)\mathbf{E}(t,\mathbf{y}+\mathcal{R}(\tau)\mathbf{u})\,d\tau\right)\cdot\nabla_v G = 0, \\ &G_{|t=0} = \frac{1}{2\pi}f_0. \end{aligned} \tag{3.28}$$

Always applying Theorem 1.3, we also deduce that the weak$-*$ limit f of (f^ϵ) is the unique solution of (1.5).

The fact that the whole sequence (f^ϵ) 2-scale converges to F and weak$-*$ converges to f is a direct consequence of the uniqueness of the solution of (3.28) and (1.5). This ends the proof of Theorem 1.1.

3.3. Link with physical models. In order to compare the model we obtain with the finite Larmor radius approximation used by physicists, we restrain to the plane orthogonal to the magnetic field. Denoting here $R(\tau)$ and $\mathcal{R}(\tau)$ there restrictions to this plan, we introduce the Larmor radius variable $\mathbf{r} = \mathbf{v}^\perp$ and the guiding center variable $\mathbf{x}_C = \mathbf{x} - \mathbf{r}$, where for any vector $\mathbf{v} = (v_2, v_3)$, $\mathbf{v}^\perp$ stand for $\mathbf{v}^\perp = (-v_3, v_2)$. In this new variables, (3.28) reads:

$$\frac{\partial f}{\partial t} - \frac{1}{2\pi}\left(\int_0^{2\pi} \mathbf{E}^\perp(t, \mathbf{x}_C + R(\tau)\mathbf{r})\, d\tau\right)\cdot \nabla_{x_C} f + \frac{1}{2\pi}\left(\int_0^{2\pi} R(-\tau)\mathbf{E}^\perp(t, \mathbf{x}_C + R(\tau)\mathbf{r})\, d\tau\right)\cdot \nabla_r f = 0. \tag{3.29}$$

Then, assuming that the distribution function is a Maxwellian distribution, i.e., $f \equiv n(\mathbf{x}_c, t)e^{-\mathbf{r}^2/(2\sigma^2)}/(2\pi\sigma)$, we integrate (3.29) with respect to $\mathbf{r}$. This procedure cancels the third term. Indeed

$$\int_0^{2\pi} R(-\tau)\mathbf{E}^\perp(t, \mathbf{x}_C + R(\tau)\mathbf{r})\, d\tau \tag{3.30}$$

depends only on $|\mathbf{r}|$ and then the integrand is odd. Then we get

$$\frac{\partial n}{\partial t} - \int_{\mathbb{R}^2_r} \frac{1}{2\pi}\left(\int_0^{2\pi} \mathbf{E}^\perp(t, \mathbf{x}_C + R(\tau)\mathbf{r})\, d\tau\right) e^{-\mathbf{r}^2/(2\sigma^2)}/(2\pi\sigma)\, d\mathbf{r}\cdot \nabla_{x_C} n = 0, \tag{3.31}$$

which is the model introduced by Hansen et al. [15].

3.4. About previous results. Notice that Theorem 1.1 of Frénod and Sonnendrücker [8], proving that the asymptotic behavior of

$$\begin{aligned} &\frac{\partial f^\epsilon}{\partial t} + \mathbf{v}\cdot\nabla_x f^\epsilon + \left(\mathbf{E} + \mathbf{v}\times\left(\mathbf{B} + \frac{\mathbf{m}}{\epsilon}\right)\right)\cdot\nabla_v f^\epsilon = 0, \\ &f^\epsilon_{|t=0} = f_0 \end{aligned} \tag{3.32}$$

is given by

$$\begin{aligned} &\frac{\partial f}{\partial t} + \mathbf{v}_\parallel\cdot\nabla_x f + (\mathbf{E}_\parallel + \mathbf{v}\times\mathbf{B}_\parallel)\cdot\nabla_v f = 0, \\ &f_{|t=0} = \frac{1}{2\pi}\int_0^{2\pi} f_0(\mathbf{x}, \mathbf{u}(\mathbf{v}, \tau))\, d\tau, \end{aligned} \tag{3.33}$$

is also a consequence of Theorem 1.3 by setting

$$\mathbf{A} = \begin{pmatrix} \mathbf{v} \\ \mathbf{E} + \mathbf{v}\times\mathbf{B} \end{pmatrix} \text{ and } \mathbf{L} = \begin{pmatrix} 0 \\ \mathbf{v}\times\mathbf{m} \end{pmatrix}. \tag{3.34}$$

This is the same for Theorem 3.2 of [8] with

$$\mathbf{A} = \begin{pmatrix} \mathbf{v} \\ \mathbf{E} + \mathbf{v} \times \mathbf{B} \end{pmatrix} \text{ and } \mathbf{L} = \begin{pmatrix} 0 \\ \mathbf{n} + \mathbf{v} \times \mathbf{m} \end{pmatrix}. \tag{3.35}$$

This theorem says that the weak$-*$ limit of the solution of

$$\begin{aligned} &\frac{\partial f^\epsilon}{\partial t} + \mathbf{v} \cdot \nabla_x f^\epsilon + \left(\left(\mathbf{E} + \frac{\mathbf{n}}{\epsilon} \right) + \mathbf{v} \times \left(\mathbf{B} + \frac{\mathbf{m}}{\epsilon} \right) \right) \cdot \nabla_v f^\epsilon = 0, \\ &f^\epsilon_{|t=0} = f_0, \end{aligned} \tag{3.36}$$

with $\mathbf{n} = \mathbf{e}_2$, satisfies

$$\begin{aligned} &\frac{\partial f}{\partial t} + \begin{pmatrix} v_1 \\ 0 \\ -1 \end{pmatrix} \cdot \nabla_x f + \left[\begin{pmatrix} E_1 - B_2 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} v_1 \\ v_2 \\ v_3 + 1 \end{pmatrix} \times \begin{pmatrix} B_1 \\ 0 \\ 0 \end{pmatrix} \right] \cdot \nabla_v f = 0, \\ &f_{|t=0} = \frac{1}{2\pi} \int_0^{2\pi} f_0(\mathbf{x}, \mathbf{u}(\mathbf{v}, \tau))\, d\tau. \end{aligned} \tag{3.37}$$

4. 2-scale limit of the 2D Vlasov–Poisson system. The aim of this section is to characterize the equation satisfied by the 2-scale limit of the sequence $(f^\epsilon, \mathbf{E}^\epsilon)$ of the Vlasov–Poisson system (1.17) For this purpose, we generalize the generic framework to the case when the operator $\mathbf{A}^\epsilon$ is oscillating. Then we apply the results obtained in this new generic framework to prove Theorem 1.4.

4.1. Generalized generic framework–proof of Theorem 1.5. We consider here

$$\begin{aligned} &\frac{\partial u^\epsilon}{\partial t} + \mathbf{A}^\epsilon \cdot \nabla_x u^\epsilon + \frac{1}{\epsilon} \mathbf{L} \cdot \nabla_x u^\epsilon = 0, \\ &u^\epsilon_{t=0} = u_0, \end{aligned} \tag{4.1}$$

where the notations are the same as for (1.10): $u^\epsilon \equiv u^\epsilon(t, \mathbf{x})$, $t \in [0, T)$, $T < \infty$; $\mathbf{x} \in \mathbb{R}^n = \mathcal{O}$, $\mathcal{Q} = [0, T) \times \mathcal{O}$. We suppose, as previously, that $\mathbf{L} \equiv M\mathbf{x} + N$, where M is a constant entry matrix satisfying $\mathrm{tr} M = 0$ and $e^{\tau M}$ is θ-periodic. The assumptions we make on $\mathbf{A}^\epsilon$ are the following: we suppose that, for all $\epsilon > 0$, $\nabla_x \cdot \mathbf{A}^\epsilon = 0$ and that, for some $q > 1$,

$$\mathbf{A}^\epsilon \text{ 2-scale converges to } \mathcal{A} \in L^\infty(0, T; L^\infty_\theta(\mathbb{R}_\tau; W^{1,q}(K))) \tag{4.2}$$

for all compact sets $K \subset \mathbb{R}^n$ and where $\mathcal{A} \equiv \mathcal{A}(t, \tau, \mathbf{x})$ is θ-periodic in τ.

The proof of Theorem 1.5 begins as the proof of Theorem 1.3 in the sense that the constraint equation and its consequences are similar. Hence relation (1.28) is obvious. In order to get the equation U_0 satisfies, we proceed as follows: we define $w^\epsilon(t, \mathbf{x}) = u^\epsilon(t, e^{\frac{t}{\epsilon} M}(\mathbf{x} - \overline{N}) + \overline{N})$, which is the function u^ϵ from which the essential oscillation is removed. This idea has also been used in Frénod and Sonnendrücker [10], Grenier [13, 14], Schochet [20], Joly, Metivier, and Rauch [16], and Colin [5]. Using the equation satisfied by w^ϵ, we show that

$$w^\epsilon \to \theta U_0 \text{ in } L^\infty(0, T; (W^{1,q}_0(K))^*), \tag{4.3}$$

where $(W_0^{1,q}(K))^*$ is the dual of $(W_0^{1,q}(K))$. This fact coupled with the assumption on $\mathbf{A}^\epsilon$ enables us to pass to the limit and find (1.29).

Under assumption (1.26) we may deduce, up to subsequences,

$$u^\epsilon \text{ 2-scale converges to } U \in L^\infty(0,T;L^\infty_\theta(\mathbb{R}_\tau;L^p(\mathcal{O}))). \tag{4.4}$$

The weak formulation of (4.1) with θ-periodic oscillating functions $(\psi)^\epsilon \equiv \psi(t,\frac{t}{\epsilon},\mathbf{x})$ writes

$$\int_{\mathcal{Q}} u^\epsilon \left(\left(\frac{\partial\psi}{\partial t}\right)^\epsilon + \frac{1}{\epsilon}\left(\frac{\partial\psi}{\partial\tau}\right)^\epsilon + \mathbf{A}^\epsilon\cdot(\nabla_x\psi)^\epsilon + \frac{1}{\epsilon}\mathbf{L}\cdot(\nabla_x\psi)^\epsilon \right) dtd\mathbf{x} = -\int_{\mathcal{O}} u_0\psi(0,0,\mathbf{x})\,d\mathbf{x}. \tag{4.5}$$

Proceeding as in subsection 3.1 we obtain that U satisfies

$$\frac{\partial U}{\partial\tau} + \mathbf{L}\cdot\nabla_x U = 0 \text{ in } \mathcal{D}'(\mathbb{R}_\tau\times\mathcal{O}) \tag{4.6}$$

and then

$$U(t,\tau,\mathbf{x}) = U_0(t, e^{-\tau M}(\mathbf{x}-\overline{N})+\overline{N}) \tag{4.7}$$

for $U_0 \equiv U_0(t,\mathbf{y}) \in L^\infty(0,T;L^2(\mathcal{O}'))$, which is (1.28) of Theorem 1.5.

Now we look for the equation U_0 satisfies. For this purpose, we define

$$w^\epsilon(t,\mathbf{y}) = u^\epsilon(t, e^{\frac{t}{\epsilon}M}(\mathbf{y}-\overline{N})+\overline{N}), \tag{4.8}$$

and we have the following lemma which characterizes the asymptotic limit of w^ϵ.

LEMMA 4.1. *The sequence (w^ϵ) satisfies*

$$w^\epsilon \to \theta U_0 \text{ in } L^\infty(0,T;(W_0^{1,q}(K))^*), \tag{4.9}$$

where U_0 is linked with the profile U by (1.28).

Proof. First, we prove that w^ϵ 2-scale converges to U_0 and w^ϵ weakly$-*$ converges to θU_0. Second, we show that $\frac{\partial w^\epsilon}{\partial t}$ is bounded in $L^\infty(0,T;(W_0^{1,r}(K))^*)$ for some $r>1$. Since w^ϵ is bounded in $L^\infty(0,T;L^p(\mathcal{O}))$, the Aubin–Lions lemma leads to the conclusion.

We take any function $\phi(t,\tau,\mathbf{y})$ regular, with compact support in t and $\mathbf{y}$ and θ-periodic with respect to τ. We have

$$\begin{aligned}\int_{\mathcal{O}'} w^\epsilon(t,\mathbf{y})\phi\left(t,\frac{t}{\epsilon},\mathbf{y}\right) dtd\mathbf{y} &= \int_{\mathcal{O}'} u^\epsilon(t,e^{\frac{t}{\epsilon}M}(\mathbf{y}-\overline{N})+\overline{N})\phi\left(t,\frac{t}{\epsilon},\mathbf{y}\right) dtd\mathbf{y}\\ &= \int_{\mathcal{O}} u^\epsilon(t,\mathbf{x})\phi\left(t,\frac{t}{\epsilon},e^{-\frac{t}{\epsilon}M}(\mathbf{x}-\overline{N})+\overline{N}\right) dtd\mathbf{x}.\end{aligned} \tag{4.10}$$

This last quantity converges to

$$\int_{\mathcal{O}}\int_0^\theta U(t,\tau,\mathbf{x})\phi(t,\tau,e^{-\tau M}(\mathbf{x}-\overline{N})+\overline{N})\,dtd\mathbf{x}d\tau = \int_{\mathcal{O}}\int_0^\theta U_0\phi\,dtd\mathbf{y}d\tau, \tag{4.11}$$

proving w^ϵ 2-scale converges to U_0. Since U_0 does not depend on τ, we immediately deduce $w^\epsilon \rightharpoonup \theta U_0$ weakly$-*$.

Now, we seek the equation w^ϵ satisfies. We have

(4.12)
$$\frac{\partial w^\epsilon}{\partial t}(t,\mathbf{y}) = \frac{\partial u^\epsilon}{\partial t}(t, e^{\frac{t}{\epsilon}M}(\mathbf{y}-\overline{N})+\overline{N}) + \frac{M}{\epsilon}e^{\frac{t}{\epsilon}M}(\mathbf{y}-\overline{N})\cdot\nabla_x u^\epsilon(t, e^{\frac{t}{\epsilon}M}(\mathbf{y}-\overline{N})+\overline{N});$$

writing this last equality in $\mathbf{y} = e^{-\frac{t}{\epsilon}M}(\mathbf{x}-\overline{N})+\overline{N}$ we obtain

$$\begin{aligned}\frac{\partial w^\epsilon}{\partial t}(t, e^{-\frac{t}{\epsilon}M}(\mathbf{x}-\overline{N})+\overline{N}) &= \frac{\partial u^\epsilon}{\partial t}(t,\mathbf{x}) + \frac{M}{\epsilon}(\mathbf{x}-\overline{N})\cdot\nabla_x u^\epsilon(t,\mathbf{x}) \\ &= \frac{\partial u^\epsilon}{\partial t}(t,\mathbf{x}) + \frac{1}{\epsilon}\mathbf{L}\cdot\nabla_x u^\epsilon(t,\mathbf{x}).\end{aligned} \tag{4.13}$$

Hence in view of the equation satisfied by u^ϵ and of

$$\nabla_y w^\epsilon(t,\mathbf{y}) = (e^{\frac{t}{\epsilon}M})^T\nabla_x u^\epsilon(t, e^{\frac{t}{\epsilon}M}(\mathbf{y}-\overline{N})+\overline{N}), \tag{4.14}$$

we obtain that w^ϵ is solution of

$$\frac{\partial w^\epsilon}{\partial t} + \mathbf{A}^\epsilon(t, e^{\frac{t}{\epsilon}M}(\mathbf{y}-\overline{N})+\overline{N})\cdot(e^{-\frac{t}{\epsilon}M})^T\nabla_y w^\epsilon = 0, \tag{4.15}$$

i.e.,

$$\frac{\partial w^\epsilon}{\partial t} + e^{-\frac{t}{\epsilon}M}\mathbf{A}^\epsilon(t, e^{\frac{t}{\epsilon}M}(\mathbf{y}-\overline{N})+\overline{N})\cdot\nabla_y w^\epsilon = 0. \tag{4.16}$$

Having (4.16) at hand we can prove that $\frac{\partial w^\epsilon}{\partial t}$ is bounded in $L^\infty(0,T;(W_0^{1,r}(K))^*)$ for some $r>1$ and any compact $K\subset\mathbb{R}^n$. It is an easy game to show

$$\nabla_y\cdot\left[e^{-\frac{t}{\epsilon}M}\mathbf{A}^\epsilon(t, e^{\frac{t}{\epsilon}M}(\mathbf{y}-\overline{N})+\overline{N})\right] = 0. \tag{4.17}$$

Hence, from (4.16) we deduce

$$\frac{\partial w^\epsilon}{\partial t} = -\nabla_y\cdot\left[e^{-\frac{t}{\epsilon}M}\mathbf{A}^\epsilon(t, e^{\frac{t}{\epsilon}M}(\mathbf{y}-\overline{N})+\overline{N})\; w^\epsilon\right], \tag{4.18}$$

and since, due to its two scale convergence, $\mathbf{A}^\epsilon$ is bounded in $W_0^{1,q}(K)$ with a bound independent on t, a Sobolev embedding theorem implies that $\mathbf{A}^\epsilon$ is bounded in $L^{q'}(K)$, where q' is defined by $\frac{1}{q'} = \mathrm{Max}(\frac{1}{q}-\frac{1}{n},0)$, and since $\tau\mapsto e^{-\tau M}$ is periodic, we deduce that $\left(e^{-\frac{t}{\epsilon}M}\mathbf{A}^\epsilon(t, e^{\frac{t}{\epsilon}M}(\mathbf{y}-\overline{N})+\overline{N})\right)$ is also bounded in $L^{q'}(K)$. Then as w^ϵ is bounded in $L^p(\mathcal{O})$, we get that $\left(e^{-\frac{t}{\epsilon}M}\mathbf{A}^\epsilon(t, e^{\frac{t}{\epsilon}M}(\mathbf{y}-\overline{N})+\overline{N})\, w^\epsilon\right)$ is bounded in $L^{r^*}(K)$ with $\frac{1}{r^*} = \frac{1}{p}+\frac{1}{q'}$. We may then conclude that

$$\frac{\partial w^\epsilon}{\partial t} \text{ is bounded in } L^\infty(0,T;(W_0^{1,r}(K))^*), \text{ with } \frac{1}{r^*}+\frac{1}{r} = 1 \tag{4.19}$$

for any compact $K\subset\mathbb{R}^n$.

In order to conclude, we treat first the case when $r^*<q^*$, where q^* is such that $\frac{1}{q^*}+\frac{1}{q} = 1$. As K is compact, we have $L^{q^*}(K)\subset L^{r^*}(K)$. Since, by considering separately the functions and their derivatives, $(W_0^{1,q}(K))^*$ and $(W_0^{1,r}(K))^*$ are, respectively, isomorphic to $(L^{q^*}(K))^{n+1}$ and $(L^{r^*}(K))^{n+1}$ we deduce, on the one hand,

$$(W_0^{1,q}(K))^* \subset (W_0^{1,r}(K))^* \tag{4.20}$$

with continuous injection. On the other hand, as a consequence of the Rellich–Kondrachov theorem, we have

$$L^p(K) \subset (W_0^{1,q}(K))^* \quad \text{compactly.} \tag{4.21}$$

Hence, applying an analogue of the Aubin–Lions lemma proved by Simon [21], we deduce

$$\mathcal{U} = \left\{ u \in L^\infty(0,T;L^p(K)), \frac{\partial u}{\partial t} \in L^\infty(0,T;(W_0^{1,r}(K))^*) \right\} \tag{4.22}$$

is compactly embedded in $L^\infty(0,T;(W_0^{1,q}(K))^*)$. Then we deduce that w^ϵ converges strongly in $L^\infty(0,T;(W_0^{1,r}(K))^*)$, giving the conclusion of the Lemma.

The case $r^* \geq q^*$ is simpler. Indeed, in this case, we directly have from (4.19)

$$\frac{\partial w^\epsilon}{\partial t} \text{ is bounded in } L^\infty(0,T;(W_0^{1,q}(K))^*), \tag{4.23}$$

yielding directly the conclusion of the lemma. ∎

Having Lemma 4.1 at hand, we want to pass to the limit in (4.16). This will give the equation for U_0. In order to realize this, we have first to prove the following.

LEMMA 4.2. *We have*

(4.24)
$\mathbf{A}^\epsilon(t, e^{\frac{t}{\epsilon}M}(\mathbf{y}-\overline{N})+\overline{N})$ *2-scale converges to* $\mathcal{A}(t,\tau,e^{\tau M}(\mathbf{y}-\overline{N})+\overline{N})$ *in* $L^\infty(0,T;L^\infty_\theta(\mathbb{R}_\tau;W_0^{1,q}(K)))$.

Proof. A direct computation gives

$$\begin{aligned}
&\int_{\mathcal{Q}'} \mathbf{A}^\epsilon(t, e^{\frac{t}{\epsilon}M}(\mathbf{y}-\overline{N})+\overline{N})\psi\left(t,\frac{t}{\epsilon},\mathbf{y}\right)\,dtd\mathbf{y} \\
&\qquad = \int_{\mathcal{Q}} \mathbf{A}^\epsilon(t,\mathbf{x})\psi\left(t,\frac{t}{\epsilon},e^{-\frac{t}{\epsilon}M}(\mathbf{x}-\overline{N})+\overline{N}\right)\,dtd\mathbf{x} \\
&\qquad \to \int_{\mathcal{Q}}\int_0^\theta \mathcal{A}(t,\tau,\mathbf{x})\psi(t,\tau,e^{-\tau M}(\mathbf{x}-\overline{N})+\overline{N}))\,dtd\mathbf{y}d\tau \\
&\qquad\qquad = \int_{\mathcal{Q}'}\int_0^\theta \mathcal{A}(t,\tau,e^{\tau M}(\mathbf{y}-\overline{N})+\overline{N}))\psi(t,\tau,\mathbf{y})\,dtd\mathbf{y}d\tau
\end{aligned} \tag{4.25}$$

for any θ-periodic test function. This proves the lemma. ∎

Now, writing a weak formulation of (4.16), we have

$$\int_{\mathcal{Q}'} w^\epsilon\left[\frac{\partial\varphi}{\partial t} + e^{-\frac{t}{\epsilon}M}\mathbf{A}^\epsilon(t,e^{\frac{t}{\epsilon}M}(\mathbf{y}-\overline{N})+\overline{N})\cdot\nabla_y\varphi\right]dtd\mathbf{y} = \int_{\mathcal{O}'} u_0\varphi(0,\cdot)\,d\mathbf{y} \tag{4.26}$$

for any $\varphi(t,\mathbf{y})$ regular and compactly supported in $\mathcal{Q}$. Let K be a compact containing the support of φ, since

(4.27)
$w^\epsilon \to \theta U_0$ in $L^\infty(0,T;(W_0^{1,q}(K))^*)$,

$\mathbf{A}^\epsilon(t, e^{\frac{t}{\epsilon}M}(\mathbf{y}-\overline{N})+\overline{N})$ 2-scale converges to $\mathcal{A}(t,e^{\tau M}(\mathbf{y}-\overline{N})+\overline{N})$ in $L^\infty(0,T;L^\infty_\theta(\mathbb{R}_\tau;W_0^{1,q}(K)))$,

we can pass to the limit in (4.26) and find

(4.28)
$$\int_{\mathcal{Q}'} \theta U_0 \left[\frac{\partial \varphi}{\partial t} + \int_0^\theta e^{-\tau M} \mathcal{A}(t, \tau, e^{\tau M}(\mathbf{y} - \overline{N}) + \overline{N}) d\tau \cdot \nabla_y \varphi \right] dt d\mathbf{y} = \int_{\mathcal{O}'} u_0 \varphi(0, \cdot)\, d\mathbf{y}.$$

Noticing at last that neither U_0 nor φ depend on τ, (4.28) is nothing but a weak formulation of (1.29), proving Theorem 1.5.

Remark. In the case when $\mathbf{A}$ does not depend on ϵ, its 2-scale limit is $\frac{\mathbf{A}}{\theta}$, so that we indeed get the same result as in Theorem 1.3.

4.2. Application to the 2D Vlasov–Poisson system—proof of Theorem 1.4. In order to deduce Theorem 1.4 from Theorem 1.5 we essentially have to show (1.20) and to pass to the limit in the Poisson equation. Indeed, once those two things are proved, the theorem follows noticing that the Vlasov equation which is the first equation of (1.17) enters the generalized generic framework with

$$\mathbf{A}^\epsilon(t, \mathbf{x}, \mathbf{v}) = \begin{pmatrix} 0 \\ \mathbf{E}^\epsilon(t, \mathbf{x}) \end{pmatrix} (\in \mathbb{R}^4) \text{ and } \mathbf{L}(t, \mathbf{x}, \mathbf{v}) = \begin{pmatrix} \mathbf{v} \\ \mathbf{v} \times \mathbf{m} \end{pmatrix} (\in \mathbb{R}^4). \tag{4.29}$$

In this case $e^{\tau M}$ becomes

$$e^{\tau M} = \begin{pmatrix} I & \mathcal{R}(\tau) \\ 0 & R(\tau) \end{pmatrix}, \tag{4.30}$$

with $\mathcal{R}(\tau)$ and $R(\tau)$ given by (1.23).

Multiplying the Vlasov equation which is the first equation of (1.17) by $(f^\epsilon)^{p-1}$ and integrating in $\mathbf{x}$ and $\mathbf{v}$ we obtain

$$\|f^\epsilon\|_{L^\infty(0,T;L^p(\mathcal{O}))} \leq C \tag{4.31}$$

for some constants C. From this estimate, we deduce the following.

LEMMA 4.3. *Under assumption* (1.19)

$$f^\epsilon \text{ 2-scale converges to } F \in L^\infty(0, T; L^\infty_{2\pi}(\mathbb{R}_\tau; L^p(\mathcal{O}))). \tag{4.32}$$

The fact that $\mathbf{E}^\epsilon$ 2-scale converges takes a bit longer to obtain. We need first to show the following lemma.

LEMMA 4.4. *Under assumption* (1.19)*, we have*

$$\|(|v|^2 f^\epsilon)\|_{L^\infty(0,T;L^1(\mathcal{O}))} \leq C \text{ and } \|\rho^\epsilon(\mathbf{x}, t)\|_{L^\infty(0,T;L^{\frac{3}{2}}(\mathbb{R}^3_x))} \leq C \tag{4.33}$$

for some constant C.

Proof. Multiplying the Vlasov equation by $|\mathbf{v}|^2$, and integrating with respect to $\mathbf{x}$ and $\mathbf{v}$, we get

$$\frac{d}{dt} \int_{\mathcal{O}} f^\epsilon |\mathbf{v}|^2 \, d\mathbf{v}\, d\mathbf{x} - 2 \int_{\mathbb{R}^2_x} \mathbf{J}^\epsilon \cdot \mathbf{E}^\epsilon \, d\mathbf{x} = 0, \tag{4.34}$$

where

$$\mathbf{J}^\epsilon(\mathbf{x}, t) = \int_{\mathbb{R}^2_v} \mathbf{v} f^\epsilon \, d\mathbf{v}. \tag{4.35}$$

Now, integrating the Vlasov equation in $\mathbf{v}$ gives the continuity equation

$$\frac{\partial \rho^\epsilon}{\partial t} + \frac{1}{\epsilon}\nabla \cdot \mathbf{J}^\epsilon = 0. \tag{4.36}$$

Using this, we obtain

$$\int_{\mathbb{R}^2_x} \mathbf{J}^\epsilon \cdot \mathbf{E}^\epsilon \, d\mathbf{x} = -\int_{\mathbb{R}^2_x} \mathbf{J}^\epsilon \cdot \nabla \phi^\epsilon \, d\mathbf{x} = \int_{\mathbb{R}^2_x} \nabla \cdot \mathbf{J}^\epsilon \, \phi^\epsilon \, d\mathbf{x} = -\epsilon \int_{\mathbb{R}^2_x} \frac{\partial \rho_\epsilon}{\partial t} \, \phi^\epsilon \, d\mathbf{x}. \tag{4.37}$$

Using now the Poisson equation, we get

$$\frac{1}{2}\frac{d}{dt}\int_{\mathbb{R}^2_x} (\nabla \phi^\epsilon)^2 \, d\mathbf{x} = -\int_{\mathbb{R}^2_x} \frac{\partial}{\partial t}\Delta \phi^\epsilon \, \phi^\epsilon \, d\mathbf{x} = \int_{\mathbb{R}^2_x} \frac{\partial \rho_\epsilon}{\partial t} \, \phi^\epsilon \, d\mathbf{x}. \tag{4.38}$$

Coupling (4.37) and (4.38) yields

$$-2\int_{\mathbb{R}^2_x} \mathbf{J}^\epsilon \cdot \mathbf{E}^\epsilon \, d\mathbf{x} = \epsilon \frac{d}{dt}\int_{\mathbb{R}^2_x} (\nabla \phi^\epsilon)^2 \, d\mathbf{x}, \tag{4.39}$$

and then (4.34) reads

$$\frac{d}{dt}\left[\int_{\mathcal{O}} f^\epsilon |\mathbf{v}|^2 \, d\mathbf{v} \, d\mathbf{x} + \epsilon \int_{\mathbb{R}^2_x} (\nabla \phi^\epsilon)^2 \, d\mathbf{x},\right] = 0, \tag{4.40}$$

and as an immediate consequence we have

$$\|(|v|^2 f^\epsilon)\|_{L^\infty(0,T;L^1(\mathcal{O}))} \le C \tag{4.41}$$

for some constant C. The first part of the lemma is then proved.

Concerning ρ^ϵ we have

$$\rho^\epsilon(\mathbf{x},t) = \int_{\mathbb{R}^2_v} f^\epsilon \, d\mathbf{v} = \int_{|v|<R} f^\epsilon \, d\mathbf{v} + \int_{|v|>R} f^\epsilon \, d\mathbf{v} \tag{4.42}$$

for any $R > 0$. Using the Cauchy–Schwartz inequality, we have

$$\int_{|v|<R} f^\epsilon \, d\mathbf{v} \le \left(\int_{|v|<R} (f^\epsilon)^2 \, d\mathbf{v}\right)^{\frac{1}{2}} \left(\int_{|v|<R} d\mathbf{v}\right)^{\frac{1}{2}} \le C_1 R \left(\int_{\mathbb{R}^2_v} (f^\epsilon)^2 \, d\mathbf{v}\right)^{\frac{1}{2}} \tag{4.43}$$

and

$$\int_{|v|>R} f^\epsilon \, d\mathbf{v} \le \int_{|v|>R} \frac{|v|^2}{R^2} f^\epsilon \, d\mathbf{v} \le \frac{1}{R^2}\int_{\mathbb{R}^2_v} |v|^2 f^\epsilon \, d\mathbf{v}. \tag{4.44}$$

Hence, we have for any $R > 0$

$$|\rho^\epsilon(\mathbf{x},t)| \le C_1 R \left(\int_{\mathbb{R}^2_v} (f^\epsilon)^2 \, d\mathbf{v}\right)^{\frac{1}{2}} + \frac{1}{R^2}\int_{\mathbb{R}^2_v} |v|^2 f^\epsilon \, d\mathbf{v}. \tag{4.45}$$

Taking the R which minimizes the right-hand side we obtain

$$|\rho^\epsilon(\mathbf{x},t)| \le C_2 \left(\int_{\mathbb{R}^2_v} (f^\epsilon)^2 \, d\mathbf{v}\right)^{\frac{1}{3}} \left(\int_{\mathbb{R}^2_v} |v|^2 f^\epsilon \, d\mathbf{v}\right)^{\frac{1}{3}} \tag{4.46}$$

and finally

$$\int_{\mathbb{R}^2_x} |\rho^\epsilon(\mathbf{x},t)|^{\frac{3}{2}}\,d\mathbf{x} \le C_3 \int_{\mathbb{R}^2_x} \left(\int_{\mathbb{R}^2_v} (f^\epsilon)^2\,d\mathbf{v}\right)^{\frac{1}{2}} \left(\int_{\mathbb{R}^2_v} |v|^2 f^\epsilon\,d\mathbf{v}\right)^{\frac{1}{2}} d\mathbf{x},$$
$$\le C_3 \left(\int_{\mathbb{R}^2_x\times\mathbb{R}^2_v} (f^\epsilon)^2\,d\mathbf{x}\,d\mathbf{v}\right)^{\frac{1}{2}} \left(\int_{\mathbb{R}^2_x\times\mathbb{R}^2_v} |v|^2 f^\epsilon\,d\mathbf{x}\,d\mathbf{v}\right)^{\frac{1}{2}},$$

thanks to the Hölder inequality. Now, knowing that the terms on the right-hand side are bounded, we have our estimate on ρ^ϵ. Hence the proof of the lemma is ended. □

As a direct consequence of Lemma 4.4, and of the regularization properties of the Laplace operator, we deduce that $\mathbf{E}^\epsilon$ is bounded in $L^\infty(0,T;W^{1,\frac{3}{2}}(\mathbb{R}^2_x))$ and the following lemma holds true.

LEMMA 4.5. *Extracting a subsequence, we have*

$$E^\epsilon \text{ 2-scale converges to } \mathcal{E} \in L^\infty(0,T;L^\infty_{2\pi}(\mathbb{R}_\tau;W^{1,\frac{3}{2}}(\mathbb{R}^2_x))). \tag{4.47}$$

Hence we proved the two facts yielding the first two equations of (1.22) with the help of Theorem 1.5.

It now remains to pass to the 2-scale limit in the Poisson equation (1.17). This can be done easily writing a weak formulation of the Poisson equation with oscillating test functions,

$$\int_{\mathbb{R}^2_x} \nabla\phi^\epsilon(t,\mathbf{x})\cdot\nabla\psi\left(t,\frac{t}{\epsilon},\mathbf{x}\right)\,dtd\mathbf{x} = \int_{\mathcal{O}} f^\epsilon(t,\mathbf{x},\mathbf{v})\psi\left(t,\frac{t}{\epsilon},\mathbf{x}\right)\,dtd\mathbf{x}d\mathbf{v}, \tag{4.48}$$

in which case we can pass to the limit and obtain, denoting Φ the 2-scale limit of ϕ^ϵ

$$\int_{\mathbb{R}^2_x}\int_0^{2\pi} \nabla\Phi\cdot\nabla\psi\,dtd\mathbf{x}d\tau = \int_{\mathcal{O}}\int_0^{2\pi} F\psi\,dtd\mathbf{x}d\mathbf{v}d\tau$$
$$= \int_{\mathcal{O}}\int_0^{2\pi} G(t,\mathbf{x}+\mathcal{R}(-\tau)\mathbf{v},R(-\tau)\mathbf{v})\psi\,dtd\mathbf{x}d\mathbf{v}d\tau,$$

which is the weak formulation of the third equation of (1.22), achieving, in view of what is said in the beginning of the subsection, the proof of Theorem 1.4.

Remark. The deduction of the equation satisfied by the weak$-*$ limit f from (1.22) is an open problem. Indeed, writing an equation for $[G(t,\mathbf{x}+\mathcal{R}(-\tau)\mathbf{v},R(-\tau)\mathbf{v})]$ from (1.22) introduces the τ-variable in the coefficients of $\nabla_x[G(t,\mathbf{x}+\mathcal{R}(-\tau)\mathbf{v},R(-\tau)\mathbf{v})]$ and $\nabla_v[G(t,\mathbf{x}+\mathcal{R}(-\tau)\mathbf{v},R(-\tau)\mathbf{v})]$. Hence we cannot proceed as in the linear case. Moreover, since those coefficients also depend on $\mathbf{x}$ and $\mathbf{v}$, the nonlocal homogenization methods (see Tartar [22, 23], Amirat, Hamdache and Ziani [3, 4] Frénod and Hamdache [7], Alexandre [1]...) do not work.

Acknowledgments. We would like to thank Pierre Bertrand for the very interesting discussions we had with him and which helped us understand the finite Larmor radius approximation. We also thank Pierre-Arnaud Raviart for his pertinent advice about this work.

REFERENCES

[1] R. Alexandre, *Some results in homogenization tackling memory effects*, Asymptot. Anal., 15 (1997), pp. 229–259.

[2] G. Allaire, *Homogeneization and Two-scale Convergence*, SIAM J. Math. Anal., 23 (1992), pp. 1482–1518.

[3] Y. Amirat, K. Hamdache, and A. Ziani, *Homogénéisation d'équations hyperboliques du premier ordre et application aux écoulements missibles en milieux poreux*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 6 (1989), pp. 397–417.

[4] Y. Amirat, K. Hamdache, and A. Ziani, *Homogenisation of parametrised families of hyperbolic problems*, Proc. Roy. Soc. Edinburgh Sect. A, 120 (1992), pp. 199–221.

[5] T. Colin, *Rigorous derivation of the non linear Schrödinger equation and Davey-Stewartson system from quadratic hyperbolic systems*, Personal communication.

[6] D. H. E. Dubin, J. A. Krommes, C. Oberman, and W. W. Lee, *Nonlinear gyrokinetic equations*, Phys. Fluids, 26 (1983), pp. 3524–3535.

[7] E. Frénod and K. Hamdache, *Homogeneization of kinetic equations with oscillating potentials*, Proc. Roy. Soc. Edinburgh Sect. A, 126 (1996), pp. 1247–1275.

[8] E. Frénod and E. Sonnendrücker, *Homogenization of the Vlasov equation and of the Vlasov-Poisson system with a strong external magnetic field*, Asymptot. Anal., 18 (1998), pp. 193–213.

[9] E. Frénod and E. Sonnendrücker, *Approximation "rayon de larmor fini" de l'équation de Vlasov*, C.R. Acad. Sci. Paris Sér. I Math., 330 (2000), pp. 421–426.

[10] E. Frénod and E. Sonnendrücker, *Long time behavior of the Vlasov equation with a strong external magnetic field*, Math. Models Methods Appl. Sci., 10 (2000), pp. 539–553.

[11] F. Golse and L. Saint-Raymond, *L'approximation centre-guide pour l'équation de Vlasov-Poisson* 2D, C. R. Acad. Sci. Paris Sér I. Math., 328 (1998), pp. 865–870.

[12] F. Golse and L. Saint-Raymond, *The Vlasov-Poisson system with strong magnetic field*, J. Math. Pures. Appl., 78 (1999), pp. 791–817.

[13] E. Grenier, *Oscillatory perturbation of the Navier-Stokes equations*, J. Maths. Pures Appl., 76 (1997), pp. 477–498.

[14] E. Grenier, *Pseudo-differential energy estimates of singular perturbations*, Comm. Pure Appl. Math., 50 (1997), pp. 821–865.

[15] F. Hansen, G. Knorr, J. Lynov, H. Pécseli, and J. Juul Rasmussen, *A numerical plasma simulation including finite Larmor radius effects to arbitrary order*, Plasma Physics and Controlled Fusion, 31 (1989), pp. 173–183.

[16] J.-L. Joly, G. Métivier, and J. Rauch, *Remarques sur l'optique géométrique non linéaire multidimensionnelle*, in Séminaire sur les Équations aux Dérivées Partielles, Exp. No. 1, 1990–1991, École Polytech., Palaiseau, 1991.

[17] G. Knorr and Pécseli, *Asymptotic state of the finite larmor radius guiding-centre plasma*, J. Plasma Physics, 41 (1989), pp. 157–170.

[18] G. Nguetseng, *A general convergence result for a functional related to the theory of homogenization*, SIAM J. Math. Anal., 20 (1989), pp. 608–623.

[19] P. Rutherford and E. Frieman, Phys. Fluids, 11 (1968), p. 569.

[20] S. Schochet, *Fast singular limit of hyperbolic pdes*, J. Differential Equations, 114 (1994), pp. 476–512.

[21] J. Simon, *Compact sets in the space $L^p(0,T;B)$*, Ann. Mat. Pura Appl. (4), 146 (1987), pp. 65–96.

[22] L. Tartar, *Nonlocal effects induced by homogenization*, Essays in Honor of Ennio De Giorgi, Vol. II, Birkhäuser, Boston, 1989.

[23] L. Tartar, *Memory effects an homogenization*, Arch. Rational Mech. Anal., (1990), pp. 121–133.

[24] P. Taylor and R. Hastie, *Stability of general plasma equilibria.* I: *Formal theory*, Plasma Physics, 10 (1968), p. 479.

SIAM J. MATH. ANAL.
Vol. 32, No. 6, pp. 1248–1264

ON THE TRACE APPROACH TO THE INVERSE SCATTERING PROBLEM IN DIMENSION ONE*

ALEXEI RYBKIN†

Abstract. We present an elementary procedure to derive certain trace relations for Schrödinger operators in dimension one relating potentials with some scattering data. In this way, we obtain some new trace type formulas as well as known ones previously studied by Gesztesy, Holden, Simon, and others, our a priori hypothesis on potentials being minimal. We also establish sharp conditions on absolute summability of the trace formulas.

Key words. inverse scattering, trace formulas, Krein's spectral shift function, Weyl's m-function

AMS subject classifications. Primary, 34L25, 34L40, 15A15; Secondary, 34B20

PII. S0036141000365620

1. Introduction. For Schrödinger operators $H = -d^2/dx^2 + v(x)$ in $L_2(\mathbb{R})$ with short-range potentials $v(x)$, the inverse scattering problem is well developed due to the Gel'fand–Levitan–Marchenko procedure for recovering $v(x)$ from certain scattering data. (Simon [24] has recently initiated a systematic generalization of the Gel'fand–Levitan–Marchenko theory to the case of arbitrary v.) However, this technique is rather involved and it is always desirable to have some explicit formulas for computing potentials in terms of some observable scattering data. One of the few such methods is the so-called trace approach to the inverse scattering problem. Some formulas serving for short-range, periodical, and other cases were obtained by Deift and Trubowitz [6] in the late 1970s and more recently by Venakides [26] and Craig [5] (see also [12] for extended literature), though they were not identified as trace relations. A consistent approach based on Krein's trace formulas was recently put forward and systematically studied by Gesztesy et al. [8, 11], Gesztesy [9], Gesztesy and Holden [10, 15], Gesztesy, Holden, and Simon [12], Gesztesy and Simon [13, 14], and Gesztesy and Makarov [16]. Their original idea is not to compare H with the free Hamiltonian $H_0 = -d^2/dx^2$ but to compare H with the associated self-adjoint Dirichlet operator H_{x_0} obtained from H by splitting it at a point $x = x_0$ by a Dirichlet boundary condition $u(x_0 \pm 0) = 0$. In this way, one obtains a pair of operators (H_{x_0}, H) whose resolvent difference is a rank one operator. Then the required trace formulas for recovering potentials result from regularizing Krein's trace formula for $\operatorname{tr}(H_{x_0} - H)$. Specifically, if $x = x_0$ is a point of Lebesgue continuity for v, then

$$v(x_0) = E + \lim_{\varepsilon \to +0} \int_E^\infty e^{-\varepsilon k} \{1 - 2\xi(x_0, k)\}\, dk = E + \lim_{z \to i\infty} \int_E^\infty \left(\frac{z}{k - z}\right)^2 \{1 - 2\xi(x_0, k)\}\, dk, \tag{1.1}$$

where $\xi(x_0, k)$ is the Krein spectral shift function of the pair (H_{x_0}, H) and $E = \inf \sigma(H)$. (See section 2 for notation and background information.)

*Received by the editors March 1, 2000; accepted for publication (in revised form) October 18, 2000; published electronically March 15, 2001.
http://www.siam.org/journals/sima/32-6/36562.html
†Department of Mathematical Sciences, University of Alaska-Fairbanks, P.O. Box 756660, Fairbanks, AK 99775-6660 (ffavr@uaf.edu).

The proof of (1.1) combines Krein's trace formula with path integral arguments and imposes very weak a priori conditions on $v(x)$: local integrability and essential boundedness from below. Note that all the relevant results by Deift, Craig, Trubowitz, Venakides, and others appear as specific cases of (1.1).

In the present paper, we shall introduce a new way of generating trace-type formulas similar to (1.1) that is not based upon Krein's trace formula but rests on the Weyl theory for second order differential equations and asymptotics of the Weyl m-function. (See also [21] for some applications to the Korteweg–de Vries equation.) In particular, we give an elementary proof of (1.1) under the only requirement that v be locally integrable, thus allowing the case $E = -\infty$. Moreover, we show that the second formula (1.1) holds even for points of jump discontinuity (with a substitution $v(x_0)$ for $\frac{1}{2}(v(x_0-0)+v(x_0+0))$). To give this topic full consideration we shall also improve the relevant results of [12] on absolute summability of trace relations (1.1) by finding sharp conditions. Some other applications of our approach are in preparation [23].

We shall also take a look at the circle of ideas of Gesztesy, Holden, and Simon from a different point of view motivated by the Lax–Phillips scattering theory in the interpretation of Pavlov [20]. To this end, instead of putting a boundary condition at a point x_0 we "cut off" the part of the potential $v(x)$ lying to the left from x_0 and compare $H = -d^2/dx^2 + v_{x_0}(x)$ with $H_0 = -d^2/dx^2$, where $v_{x_0}(x) = 0, x < x_0; v_{x_0}(x) = v(x), x \geq x_0$. The scattering matrix for (H, H_0) does not in general exist, but one of its elements does, namely, the reflection coefficients $R(x_0,k)$ from the left defined in a certain way. The function $R(x_0,k)$ is analytic and contractive in the upper half plane $\mathbb{C}_+$, and its complex zeros $\{z_n(x_0)\}$ make clear physical sense of resonance states. We shall provide a simple procedure of recovering potentials in terms of the reflection coefficient.

The structure of the present paper is as follows. In section 2 we agree upon the terminology and provide the relevant background material. Section 3 is the central analytic part: we state and prove some propositions which will play a crucial role in our approach. Section 4 forms the main body of this text. Here we present the abovementioned way of representation of potentials via the reflection coefficient. Here we also improve one result of Klibanov and Sacks [18] on asymptotics of the reflection coefficient on the real line. Section 5 is related to section 4—it utilizes some of the results of the previous section to make some curious conclusions on resonances. Section 6 is also central. Here we demonstrate how our approach works in settings considered by others. We give simple proofs and extend already known results on trace formula (1.1) and its analog for other than Dirichlet boundary conditions. Finally, in section 7 we consider conditions providing absolute summability of the trace formula (1.1). We prove sharp results for cascade type potentials considered in [12].

2. Notation and preliminaries. Notation. We will follow standard notation: $\mathbb{R}_\pm = (\pm\infty, 0), \mathbb{C}_\pm = \{z \in \mathbb{C} : \pm \operatorname{Im} \geq 0\}$. Scripts z and k will always stand, respectively, for a complex variable and a real one; $z \to i\infty$ will mean $|z| \to \infty$ inside any cone $0 < \varepsilon \leq \arg z \leq \pi - \varepsilon$. We will also use standard spaces $(1 \leq p < \infty)$

$$L_p(\Delta, d\mu) = \left\{ f : \|f\|_p \equiv \left(\int_\Delta |f(x)|^p \, d\mu \right)^{1/p} < \infty \right\}, \quad L_p(\Delta, dx) \equiv L_p(\Delta),$$

$$L_\infty(\Delta) = \{ f : \|f\|_\infty \equiv \operatorname{ess\,sup} |f(x)| < \infty \}, \quad L_{p,\mathrm{loc}} = \{ \cap L_p(\Delta) : \Delta \text{ is compact} \}.$$

$W_m^p(\Delta)$ denotes the usual Sobolev space whose elements have up to m distributional derivatives in $L_p(\Delta)$; $\sigma_d(H), \sigma_{sc}(H)$, and $\sigma_{ac}(H)$ are, respectively, the discrete, singular continuous, and absolutely continuous spectrums of a self-adjoint operator H.

Lebesgue points. A point $x \in \Delta$ is called a right (left) Lebesgue continuity point of a function $f(x) \in L_1(\Delta)$ if (see, e.g., [7])

$$\int_0^h |f(x \pm t) - f(x)|\, dt = o(h), \quad h \to +0.$$

If also

$$\int_{-h}^h |f(x+t) - f(x)|\, dt = o(h), \quad h \to +0,$$

x is called a Lebesgue continuity point. A right (left) Lebesgue point may coincide with the left (right) endpoint of Δ as opposed to a Lebesgue point which by definition must be an inner point of Δ.

LEMMA 2.1. *Let $\phi(x), f(x) \in L_1(\Delta)$, $\|\phi\|_1 = 1$, and $\psi(x) = \sup\{|\phi(t)| : |t| \geq |x|\} \in L_1(\Delta)$. Then*

$$f(x) = \lim_{h \to 0} \frac{1}{h} \int_\Delta \phi\left(\frac{x-t}{h}\right) f(t)\, dt$$

for every Lebesgue point x of $f(x)$.

Herglotz functions. An analytic function $f(z)$ is called Herglotz if $f(z)$ maps $\mathbb{C}_+$ into $\mathbb{C}_+$ (see, e.g., [1]).

PROPOSITION 2.2. *Let $f(z)$ be a Herglotz function. Then there exist a positive measure $d\rho$ on $\mathbb{R}$ and a real-valued ξ such that*

$$f(z) = a + bz + \int_{\mathbb{R}} \left(\frac{1}{k-z} - \frac{k}{1+k^2}\right) d\rho(k) \tag{2.1}$$

$$= \exp\left\{c + \int_{\mathbb{R}} \left(\frac{1}{k-z} - \frac{k}{1+k^2}\right) \xi(k) dk\right\}, \tag{2.2}$$

where

$$a = \operatorname{Re} f(i), \quad b \geq 0, \quad c = \operatorname{Re} \ln f(i),$$
$$\int_{\mathbb{R}} \frac{d\rho(k)}{1+k^2} < \infty, \quad \int_{\mathbb{R}} \frac{\xi(k) dk}{1+k^2} < \infty, \quad 0 \leq \xi(k) \leq 1.$$

The functions $\rho(\Delta), \xi(k)$ can be computed by

$$\rho(\Delta) = \lim_{\varepsilon \to o+} \frac{1}{\pi} \int_\Delta \operatorname{Im} f(k + i\varepsilon)\, dk, \quad \xi(k) = \frac{1}{\pi} \lim_{\varepsilon \to o+} \arg f(k + i\varepsilon).$$

Schrödinger operators. Given a real-valued potential $v \in L_{1,\mathrm{loc}}$ we introduce the differential expression

$$l = -\frac{d^2}{dx^2} + v(x), \quad x \in \mathbb{R}.$$

We associate with l several self-adjoint operators:

$$Hu = lu, \quad u \in \mathrm{Dom}\,(H) = \{f \in L_2(\mathbb{R}) : lf \in L_2(\mathbb{R})\}, \tag{2.3}$$

$$\begin{aligned} H_{x_0}^{\pm} u &= lu, \quad u \in \mathrm{Dom}\left(H_{x_0}^{\pm}\right) \\ &= \{f \in L_2(x_0, \pm\infty) : lf \in L_2(x_0, \pm\infty), f(x_0 \pm 0) = 0\}, \end{aligned} \tag{2.4}$$

$$H_{x_0} = H_{x_0}^{+} \oplus H_{x_0}^{-}. \tag{2.5}$$

A more detailed description of the domains in (2.3)–(2.5) can be found, e.g., in [19]. If a potential v is in the so-called limit circle case at $\pm\infty$, then operators in (2.3)–(2.5) are no longer well defined. In this case we have to impose some self-adjoint boundary conditions at $\pm\infty$. Throughout the paper we include the limit circle case by assuming a suitable boundary condition at $\pm\infty$. For simplicity, we decided to retain the same notation for operators (2.3)–(2.5) in both cases.

Weyl's *m*-function. Let $\theta(x,z), \varphi(x,z)$ be the solutions of

$$-u'' + v(x)\,u = zu, \quad x \in \mathbb{R}_{\pm}, \quad z \in \mathbb{C}_{+}, \tag{2.6}$$

satisfying the following boundary conditions:

$$\theta(\pm 0, z) = 1, \quad \theta'(\pm 0, z) = 0; \quad \varphi(\pm 0, z) = 0, \quad \varphi'(\pm 0, z) = 1.$$

According to the Weyl theory (see, e.g., [19, 25]), there exists a unique function $m_{\pm}(z)$ analytic in $\mathbb{C} \setminus \mathbb{R}$ with $\pm \operatorname{Im} m_{\pm}(z) \geq 0, z \in \mathbb{C}_{\pm}$, such that

$$\psi_{\pm}(x,z) = \theta(x,z) + m_{\pm}(z)\,\varphi(x,z) \in L_2(\mathbb{R}_{\pm}). \tag{2.7}$$

This function $m_{\pm}$ is called the Weyl–Titchmarsh m-function (or the Weyl m-function) associated with $H_0^{\pm}$. Sometimes for simplicity we will write $m_+ = m$. For solutions $\theta(x,z), \varphi(x,z)$ we have

$$\theta(x,z) = \cos\sqrt{z}x + \frac{1}{\sqrt{z}} \int_0^x \sin\sqrt{z}\,(x-y)\,v(y)\,\theta(y,z)\,dy, \tag{2.8}$$

$$\varphi(x,z) = \frac{\sin\sqrt{z}x}{\sqrt{z}} + \frac{1}{\sqrt{z}} \int_0^x \sin\sqrt{z}\,(x-y)\,v(y)\,\varphi(y,z)\,dy,$$

$$\left|e^{i\sqrt{z}x}\theta(x,z)\right| \leq \exp\left\{\frac{1}{|\sqrt{z}|}\int_0^x |v(y)|\,dy\right\}, \tag{2.9}$$

$$\left|e^{i\sqrt{z}x}\varphi(x,z)\right| \leq \frac{1}{|\sqrt{z}|}\exp\left\{\frac{1}{|\sqrt{z}|}\int_0^x |v(y)|\,dy\right\}.$$

Recall that if v is in the limit circle case at $\pm\infty$, then any solution of (2.6) is in $L_2(\mathbb{R}_{\pm})$ and $m_{\pm}$ is not defined uniquely, and we then follow the procedure outlined above.

The next proposition is due to Atkinson [2].

PROPOSITION 2.3. *Let $v(x) \in L_1(0, \pm\delta)$ for some $\delta > 0$; then in any cone $0 < \varepsilon < \arg z < \pi - \varepsilon$*

$$m_{\pm}(z) = \pm i\sqrt{z} - \int_0^{\pm\delta} e^{\pm 2i\sqrt{z}x} v(x)\,dx + o\left(1/\sqrt{z}\right), \quad z \to \infty. \tag{2.10}$$

Note that further terms in asymptotics (2.10) (under the same conditions) were derived in [17].

Besides the Weyl m-function $m_\pm(z)$ of the Dirichlet problem we will also consider Weyl m-functions $m_{\pm,h}(z)$ corresponding to other conditions $u'(\pm 0)+hu(\pm 0)=0$, $h\in\mathbb{R}$. Weyl m-functions $m_{\pm,h}$ are related to $m_\pm$ by the formula

$$m_{\pm,h}(z)=\frac{hm_\pm(z)-1}{m_\pm(z)+h}. \tag{2.11}$$

The reflection coefficient. Let $v_0(x)=0, x<0$; $v_0(x)=v(x), x\geq 0$. Consider the equation

$$-u''+v_0(x)u=zu,\quad x\in\mathbb{R},\quad z\in\mathbb{C}_+,$$

and introduce its solution $u(x,z)$ such that

$$\begin{aligned} &u(x,z)=e^{i\sqrt{z}x}+R(z)e^{-i\sqrt{z}x},\quad x<0,\\ &u(x,z)=c(z)\psi_+(x,z),\quad x\geq 0, \end{aligned}$$

where $\sqrt{z}$ is chosen so that $\operatorname{Im}\sqrt{z}\geq 0$. The requirement of continuity of $u(x,z)$ and $u'(x,z)$ at $x=0$ implies

$$R(z)=-\frac{m_+(z)-i\sqrt{z}}{m_+(z)+i\sqrt{z}}. \tag{2.12}$$

Due to general properties of $m_+(z)$, $R(z)$ is analytic and contractive $|R(z)|\leq 1$ in $\mathbb{C}_+$, and therefore has boundary values a.e. on $\mathbb{R}$ and admits the canonical representation (see, e.g., [7])

$$\begin{aligned} &R(z)=e^{ic}B(z)S_i(z)S_e(z),\quad z\in\mathbb{C}_+,\\ &B(z)=\prod_n\left\{\frac{1+\overline{z}_n^2}{|1+z_n^2|}\frac{z-z_n}{z-\overline{z}_n}\right\},\\ &S(z)=\exp\frac{i}{\pi}\left\{\int_{\mathbb{R}}\frac{1+zk}{k-z}\frac{\log|R(k)|\,dk+d\mu(k)}{1+k^2}+\pi az\right\}, \end{aligned} \tag{2.13}$$

where $a\geq 0, c\in\mathbb{R}$, and $\mu(k)$ is a nonnegative singular measure.

The Weyl m-function has no clear physical sense, but the function $R(z)$, following Pavlov [20], can be interpreted as the reflection coefficient of plane waves $e^{i\sqrt{z}x}$ with momentum $\sqrt{z}, (\operatorname{Im}\sqrt{z}=0)$ coming from $-\infty$ and reflected by the potential $v_0(x)$. The fact that in general $\sigma_{ac}(H)=\operatorname{clos}\{k\in\mathbb{R}:\operatorname{Im}m(k+i0)>0\}\neq\emptyset$ (therefore, $|R(k)|\leq 1$ on $\sigma_{ac}(H)$) means that the waves $e^{i\sqrt{z}x}, z\in\sigma_{ac}(H)$, are not completely reflected by $v_0(x)$ and can also propagate on $(0,\infty)$. The zeros $\{z_n\}$ of $R(z)$ in $\mathbb{C}_+$ are related to so-called resonances (surface states) by $\{k_n\}$ by $k_n=\sqrt{z_n}, \operatorname{Im}\sqrt{z_n}>0$. Note that resonances are observable physical quantities, $\left(\operatorname{Im}\sqrt{z_n}\right)^{-1}$ being interpreted as lifetimes. Note also that $\{z_n\}$, being zeros of analytic contractive function, are subject to the Blaschke condition

$$\sum_n\frac{\operatorname{Im}z_n}{|z_n+i|^2}<\infty. \tag{2.14}$$

Krein's spectral shift function. Let (H, H_0) be an abstract pair of resolvent comparable Hilbert space operators (i.e., $(H - zI)^{-1} - (H_0 - zI)^{-1}$ is of trace class for a complex z). Then (see, e.g., [3]) there exists a unique (up to an additive constant) real-valued function $\xi(k) \in L_1\left(\mathbb{R}, \frac{dk}{1+k^2}\right)$ such that the Krein trace formula

$$\operatorname{tr}\{\varphi(H) - \varphi(H_0)\} = \int_{\mathbb{R}} \varphi'(k)\, \xi(k)\, dk \tag{2.15}$$

holds; φ is an arbitrary function of some suitable class. The function $\xi(k)$ is called the spectral shift function of the pair (H, H_0), and for almost all $k \in \sigma_{ac}(H_0)$ the function $\xi(k)$ is related to the scattering matrix $S(k)$ of (H, H_0) by the Birman–Krein formula

$$\det S(k) = e^{-2\pi i \xi(k)}.$$

Note that as opposed to the scattering matrix, the spectral shift function is defined on the whole line.

3. Auxiliary results. The following lemma is a well-known fact on differentiation of asymptotic expansions of analytic functions.

LEMMA 3.1. *Let $f(z)$ be an analytic function in $\mathbb{C}_+$. If $f(z)$ admits the representation*

$$f(z) = a_0 + \frac{a_1}{z} + o\left(\frac{1}{z}\right)$$

for $z \to \infty$ inside a cone $\Gamma_\varepsilon = \{z \in \mathbb{C}_+ : 0 < \varepsilon \le \arg z \le \pi - \varepsilon\}$, then

$$a_1 = -\lim z^2 f'(z), \quad z \to \infty, \quad z \in \Gamma_\varepsilon. \tag{3.1}$$

Proof. Change z for $1/z$. Then $\Gamma_\varepsilon \to \overline{\Gamma}_\varepsilon = \{z \in \mathbb{C}_- : \overline{z} \in \Gamma_\varepsilon\}$ and

$$f(1/z) = a_0 + a_1 z + o(z). \tag{3.2}$$

Fix $z \in \overline{\Gamma}_\varepsilon$, and let $C_r(z) = \{\lambda \in \mathbb{C}_- : |\lambda - z| = r\}$ be a circle with radius $r = |z| \sin \varepsilon/2$. It follows from (3.2) that

$$\frac{1}{2\pi i} \int_{C_r(z)} \frac{f(\lambda)\, d\lambda}{(\lambda - z)^2} = \sum_{m=0}^{1} a_m \frac{1}{2\pi i} \int_{C_r(z)} \frac{(\lambda - z_0)^m\, d\lambda}{(\lambda - z)^2} + \widetilde{o}(z), \tag{3.3}$$

where for the remainder $\widetilde{o}(z)$ we have

$$\begin{aligned} |\widetilde{o}(z)| &\le r^{-1} \max_{\lambda \in C_r(z)} |\lambda|\, o(1) \\ &= \frac{|z| + r}{r} \cdot o(1) = \frac{1 + \sin \varepsilon}{\sin \varepsilon} \cdot o(1). \end{aligned}$$

Therefore, $\widetilde{o}(z) \to 0$ as $z \to \infty, z \in \overline{\Gamma}_{\varepsilon/2}$, and hence by the Cauchy theorem, (3.3) implies

$$\frac{d}{dz} f(1/z) = a_1 + \widetilde{o}(z) \to a_1, \text{ as } z \to 0, \quad z \in \overline{\Gamma}_{\varepsilon/2},$$

which implies (3.1) by substituting $1/z$ back for z. □

LEMMA 3.2. *Let $f(x) \in L_1(0,\delta)$ for some $\delta > 0$, and let $x = 0$ be a right Lebesgue point of $f(x)$; then*

$$\int_0^\delta e^{izx} f(x)\,dx = \frac{if(0)}{z} + o(1/z), \quad z \to \infty, \quad \varepsilon < \arg z < \frac{\pi}{2} - \varepsilon.$$

Proof. We have

$$\int_0^\delta e^{izx} f(x)\,dx = f(0)\int_0^\delta e^{izx}dx + \int_0^\delta e^{izx}(f(x) - f(0))\,dx = f(0)\,\frac{e^{iz\delta}-1}{iz} + \widetilde{o}(1/z),$$

where $\widetilde{o}(1/z) = \int_0^\delta e^{izx}(f(x) - f(0))\,dx$. It follows from $\varepsilon < \arg z < \frac{\pi}{2} - \varepsilon$ that $|z| \le (1+\cot\varepsilon)\operatorname{Im} z$, and hence one gets $\left|f(0)\,e^{iz\delta}\right| \to 0, \operatorname{Im} z \to \infty$, and

$$\begin{aligned}|z\widetilde{o}(1/z)| &\le |z|\int_0^\delta e^{-\operatorname{Im} z\cdot x}\,|f(x) - f(0)|\,dx \\ &\le (1+\cot\varepsilon)\operatorname{Im} z\int_0^\delta e^{-\operatorname{Im} z\cdot x}\,|f(x) - f(0)|\,dx = o(1), \quad \operatorname{Im} z \to \infty,\end{aligned}$$

since the functions $\psi(x) = \sup\left\{e^{-|t|} : |t| \ge |x|\right\}$ and $|f(x) - f(0)|$ satisfy the conditions of Lemma 2.1 with $h = (\operatorname{Im} z)^{-1}$. □

PROPOSITION 3.3. *Suppose $v(x)$ is a real-valued function defined on $\mathbb{R}$ and such that*

$$\exists v_\pm \in \mathbb{R} : v(x) - v_\pm \in L_1(\mathbb{R}_\pm), \quad \int_{\mathbb{R}_\pm} (v(x) - v_\pm)\cos sx dx \in L_1(\mathbb{R}_+). \tag{3.4}$$

Set $\Omega_\pm(\lambda) \equiv \frac{m_\pm(\lambda+i0) \mp i\sqrt{\lambda - v_\pm}}{\pm i\sqrt{\lambda}}$; then for some $a > 0$

$$\operatorname{Im}\Omega_\pm, \quad \operatorname{Im}\Omega_\pm^2 \in L_1(a,\infty).$$

Proof. Consider only the right half-line. Observe once that $m(z) = \widetilde{m}(z-c)$, where $\widetilde{m}(z)$ is the Weyl m-function of H_0^+ with $v(x) = v(x) - v_+$, and hence we may simply set $v_+ = 0$.

Since now $v \in L_1(\mathbb{R}_+)$, we may use the well-known representation of the Weyl m-function [25]

$$m(z) = i\sqrt{z}\,\frac{1 - \int_0^\infty \frac{e^{i\sqrt{z}y}}{i\sqrt{z}} v(y)\,\theta(y,z)\,dy}{1 + \int_0^\infty e^{i\sqrt{z}y} v(y)\,\varphi(y,z)\,dy} \equiv i\sqrt{z}\,\frac{1 - \frac{1}{i\sqrt{z}}A_\theta(z)}{1 + \frac{1}{i\sqrt{z}}A_\varphi(z)}, \quad \operatorname{Im}\sqrt{z} \ge 0. \tag{3.5}$$

It follows from (3.5) that uniformly in $\operatorname{Im}\sqrt{z} \ge 0, |z| \to \infty$,

$$\frac{m(z) - i\sqrt{z}}{i\sqrt{z}} = -\frac{A_\theta(z) + A_\varphi(z)}{i\sqrt{z}} + \frac{(A_\theta(z) + A_\varphi(z))\,A_\varphi(z)}{(i\sqrt{z})^2} + O\left(\frac{e^{-3\operatorname{Im}\sqrt{z}}}{|z|^{3/2}}\right). \tag{3.6}$$

Set $z = \lambda + i\varepsilon$. Making use of (2.9), by Lebesgue's dominated convergence theorem as $\varepsilon \to 0$ we get

$$\begin{aligned}A_\theta \equiv A_\theta(\lambda) &= \int_0^\infty e^{i\sqrt{\lambda}x} v(x)\,\theta(x,\lambda+i0)\,dx, \\ A_\varphi \equiv A_\varphi(\lambda) &= i\sqrt{\lambda}\int_0^\infty e^{i\sqrt{\lambda}x} v(x)\,\varphi(x,\lambda+i0)\,dx.\end{aligned}$$

Set $\Omega = \Omega_+$. Now (3.6) yields ($\Omega = \Omega_+$)

$$\Omega(\lambda) = \frac{m(\lambda) - i\sqrt{\lambda}}{i\sqrt{\lambda}} = -\frac{A_\theta + A_\varphi}{i\sqrt{\lambda}} - \frac{(A_\theta + A_\varphi) A_\varphi}{\lambda} + O\left(\frac{1}{\lambda^{3/2}}\right), \quad \lambda \to \infty. \tag{3.7}$$

Let us compute A_θ, A_φ. Applying (2.8) twice and taking into account (2.9) one can get

$$\begin{aligned} A_\theta &= \int_0^\infty e^{i\sqrt{\lambda}x} v(x) \left\{ \cos\sqrt{\lambda}x + \frac{1}{k}\int_0^x \sin\sqrt{\lambda}(x-y)\, v(y)\, \theta(y, \lambda + i0)\, dy \right\} dx \\ &= \int_0^\infty e^{i\sqrt{\lambda}x} v(x) \cos\sqrt{\lambda}x \cdot dx \\ &+ \frac{1}{\sqrt{\lambda}} \int_0^\infty e^{i\sqrt{\lambda}x} v(x) \left\{ \int_0^x \sin\sqrt{\lambda}(x-y)\, v(y) \cos\sqrt{\lambda}y \cdot dy \right\} dx + O\left(\frac{1}{\lambda}\right). \end{aligned}$$

In the same way one obtains

$$\begin{aligned} A_\phi &= i\int_0^\infty e^{i\sqrt{\lambda}x} v(x) \sin\sqrt{\lambda}x \cdot dx \\ &+ \frac{i}{k} \int_0^\infty e^{i\sqrt{\lambda}x} v(x) \left\{ \int_0^x \sin\sqrt{\lambda}(x-y)\, v(y) \sin\sqrt{\lambda}y \cdot dy \right\} dx + O\left(\frac{1}{\lambda}\right), \end{aligned} \tag{3.8}$$

and hence

$$\begin{aligned} A_\theta + A_\varphi &= \int_0^\infty e^{2i\sqrt{\lambda}x} v(x)\, dx \\ &+ \frac{1}{k} \int_0^\infty e^{2i\sqrt{\lambda}x} v(x) \left\{ \int_0^x \sin\sqrt{\lambda}(x-y)\, v(y)\, e^{i\sqrt{\lambda}y} dy \right\} dx + O\left(\frac{1}{\lambda}\right). \end{aligned} \tag{3.9}$$

The second integral in (3.9) can be transformed into

$$\begin{aligned} &\frac{1}{2i\sqrt{\lambda}} \left\{ \int_0^\infty e^{i\sqrt{\lambda}x} v(x) \left(\int_0^x v(y)\, dy \right) dx - \int_0^\infty v(x) \left(\int_0^x e^{2i\sqrt{\lambda}y} v(y)\, dy \right) dx \right\} \\ &= \frac{1}{2i\sqrt{\lambda}} \left\{ 2\int_0^\infty e^{2i\sqrt{\lambda}x} v(x) \left(\int_0^x v(y)\, dy \right) dx - \int_0^\infty v(x)\, dx \cdot \int_0^\infty e^{2i\sqrt{\lambda}x} v(x)\, dx \right\}, \end{aligned}$$

and finally we have

$$A_\theta + A_\varphi = F\left(2\sqrt{\lambda}\right) + \frac{1}{2i\sqrt{\lambda}} \left\{ 2F_1\left(2\sqrt{\lambda}\right) - \int_0^\infty v(x)\, dx \cdot F\left(2\sqrt{\lambda}\right) \right\} + O\left(\frac{1}{\lambda}\right), \tag{3.10}$$

where

$$F(s) \equiv \int_0^\infty e^{isx} v(x)\, dx, \tag{3.11}$$

$$F_1(s) \equiv \int_0^\infty e^{isx} V(x)\, dx, \quad V(x) = v(x) \int_0^x v(y)\, dy. \tag{3.12}$$

Equation (3.8) yields

$$\begin{aligned} A_\varphi &= i\int_0^\infty e^{i\sqrt{\lambda}x} v(x) \sin\sqrt{\lambda}x \cdot dx + O\left(\frac{1}{\sqrt{\lambda}}\right) \\ &= \frac{1}{2}F\left(2\sqrt{\lambda}\right) - \frac{1}{2}\int_0^\infty v(x)\,dx + O\left(\frac{1}{\sqrt{\lambda}}\right). \end{aligned} \tag{3.13}$$

Combining (3.7), (3.10), and (3.13) one obtains

$$\Omega(\lambda) = -\frac{F\left(2\sqrt{\lambda}\right)}{i\sqrt{\lambda}} - \frac{F^2\left(2\sqrt{\lambda}\right) - 2F_1\left(2\sqrt{\lambda}\right)}{2\lambda} + O\left(\frac{1}{\lambda^{3/2}}\right), \tag{3.14}$$

$$\Omega^2(\lambda) = \frac{F^2\left(2\sqrt{\lambda}\right)}{\lambda} + O\left(1/\lambda^{3/2}\right),$$

and hence

$$\operatorname{Im}\Omega(\lambda) = \frac{\widehat{v}_c\left(2\sqrt{\lambda}\right)}{\sqrt{\lambda}} + \frac{\left(\widehat{V}_s - \widehat{v}_c\widehat{v}_s\right)\left(2\sqrt{\lambda}\right)}{\lambda} + O\left(\frac{1}{\lambda^{3/2}}\right), \tag{3.15}$$

$$\operatorname{Im}\Omega^2(\lambda) = \frac{2\widehat{v}_c\left(2\sqrt{\lambda}\right)\widehat{v}_s\left(2\sqrt{\lambda}\right)}{\lambda} + O\left(\frac{1}{\lambda^{3/2}}\right),$$

where $\widehat{f}_c(s)$ ($\widehat{f}_s(s)$) stands for $\cos$ ($\sin$) $-$Fourier transform of $f(x)$. The first term in (3.15) is in $L_1(\mathbb{R}_+)$ since

$$\left\|\lambda^{-1/2}\widehat{v}_c\left(2\sqrt{\lambda}\right)\right\|_1 = \|\widehat{v}_c\|_1.$$

Further, it is well known [27] that if the Fourier transform of a summable function is summable, then the function is continuous (together with its Fourier transform). By our condition $\widehat{v}_c \in L_1(\mathbb{R}_+)$, and hence $v \in C(\mathbb{R}_+)$. The latter implies $V(x) = v(x)\int_0^x v(y)\,dy \in C(\mathbb{R}_+)$, and hence $v, V \in L_2(\mathbb{R}_+)$ since $C(\mathbb{R}_+)\cap L_1(\mathbb{R}_+) \subset L_2(\mathbb{R}_+)$. However, by [27] the Fourier transform of a function from $L_2(\mathbb{R}_+)$ is again in $L_2(\mathbb{R}_+)$, and thus $\widehat{v}_c, \widehat{v}_s, \widehat{V}_c \in L_2(\mathbb{R}_+)$. Therefore,

$$\left\|\lambda^{-1}\widehat{v}_c\left(2\sqrt{\lambda}\right)\widehat{v}_s\left(2\sqrt{\lambda}\right)\right\|_1 \le \|\widehat{v}_c\|_2\,\|\widehat{v}_s\|_2 < \infty,$$

$$\left\|\lambda^{-1}\widehat{V}_s\left(2\sqrt{\lambda}\right)\right\|_1 = 2\left\|\lambda^{-1}\widehat{V}_s(\lambda)\right\|_1 \le C\left\|\widehat{V}_s\right\|_2 < \infty$$

with some constant C dependent on the choice of $a > 0$ in $L_2(a,\infty)$. Now from (3.15) we conclude that $\operatorname{Im}\Omega, \operatorname{Im}\Omega^2 \in L_1(a,\infty)$. □

Actually, as a byproduct, we obtained the asymptotics of the Weyl m-function along the real line (which is likely known to the specialists).

COROLLARY 3.4. *Let $v(x) \in L_1(\mathbb{R}_+)$. Then for $k \to \infty$ along the real line*

$$m(\lambda) = i\sqrt{\lambda} - F\left(2\sqrt{\lambda}\right) + \frac{1}{2i\sqrt{\lambda}}\left\{F^2\left(2\sqrt{\lambda}\right) - 2F_1\left(2\sqrt{\lambda}\right)\right\} + O\left(\frac{1}{\lambda}\right), \tag{3.16}$$

where F, F_1 are given by (3.11), (3.12).

4. Scattering on truncated potentials. Throughout this section we assume potentials to be real-valued, integrable in some neighborhood of a point $x = x_0$, and otherwise arbitrary. For simplicity set $x_0 = 0$.

THEOREM 4.1. *Suppose $v(x)$ is real-valued and belongs to $L_1(0,\delta)$ with some $\delta > 0$. If 0 is a right Lebesgue point of v and $R(z)$ is the reflection coefficient defined by (2.12), then $v(0)$ can be computed by any of the following formulas:*

$$v(0) = 4 \lim_{z\to i\infty} zR(z), \tag{4.1}$$

$$v(0) = -\frac{4}{\pi} \lim_{z\to i\infty} \int_{-\infty}^{\infty} \left(\frac{z}{k-z}\right)^2 \arg(1+R(k))\, dk, \tag{4.2}$$

$$v(0) = \frac{4i}{\pi} \lim_{z\to i\infty} \int_{-\infty}^{\infty} \left(\frac{z}{k-z}\right)^2 \operatorname{Re} R(k)\, dk. \tag{4.3}$$

Proof. It immediately follows from (2.10) and (2.12) that

$$R(z) = \frac{1}{2i\sqrt{z}} \int_0^{\delta} e^{2i\sqrt{z}x} v(x)\, dx + o(1/z), \quad z \to i\infty.$$

Then by Lemma 3.2

$$R(z) = \frac{1}{2i\sqrt{z}} \left(\frac{iv(0)}{2\sqrt{z}} + o\left(1/\sqrt{z}\right)\right) = \frac{v(0)}{4z} + o(1/z), \quad z \to \infty,$$

and we arrive at (4.1).

Now consider the function $i(1+R(z))$. Since $|R(z)| \le 1, \operatorname{Im} z \ge 0$, it is Herglotz and hence admits representation (2.2):

$$\begin{aligned} i(1+R(z)) &= \exp\left\{c + \int_{\mathbb{R}} \left(\frac{1}{k-z} - \frac{k}{1+k^2}\right) \frac{\arg(1+R(k))}{\pi} dk\right\} \\ &= i\left(1 + \frac{v(0)}{4z} + o\left(\frac{1}{z}\right)\right). \end{aligned}$$

Taking the logarithm of both parts we get

$$c + \int_{\mathbb{R}} \left(\frac{1}{k-z} - \frac{k}{1+k^2}\right) \frac{\arg(1+R(k))}{\pi} dk = i\pi/2 + \frac{v(0)}{4z} + o(1/z), \quad z \to i\infty.$$

Applying Lemma 3.1 then yields (4.2).

Using the very same arguments and (2.1), one easily obtains

$$v(0) = \frac{4i}{\pi} \lim_{z\to i\infty} \int_{-\infty}^{\infty} \left(\frac{z}{k-z}\right)^2 (1 + \operatorname{Re} R(k))\, dk.$$

Since by Cauchy's theorem $\int_{-\infty}^{\infty} \left(\frac{z}{k-z}\right)^2 dk = 0$, we arrive at (4.3). ∎

Remark 1. If $v(x) \ge c > 0$, then (4.2) reads

$$v(0) = -\frac{4}{\pi} \lim_{z\to i\infty} \int_{E}^{\infty} \left(\frac{z}{k-z}\right)^2 \arg(1+R(k))\, dk,$$

where $E = \inf \sigma\left(H_0^+\right)$. Indeed, in this case the spectrum of H_0^+ is bounded below. Due to (2.12) $R(k)$ is real-valued for $k < E$ and $\arg(1+R(k)) = 0$, $k < E$.

The limit in (4.2) cannot be removed in general. However, in a particular case of cascade-type potentials one can pass to the limit under the integral sign. See the following theorem.

THEOREM 4.2. *Suppose $v(x)$ is subject to*

$$\exists c : v(x) - c \in L_1(\mathbb{R}_+), \quad \int_0^\infty (v(x) - c)\cos kx \cdot dx \in L_1(\mathbb{R}_+). \tag{4.4}$$

Then

$$v(0) = -\frac{4}{\pi}\int_E^\infty \arg(1 + R(k))\, dk, \quad E = \inf\sigma\left(H_0^+\right). \tag{4.5}$$

Proof. First of all, as in the proof of Proposition 3.3 we may set $c = 0$. From (2.12) one derives

$$1 + R(k) = \frac{2i\sqrt{k}}{m(k) + i\sqrt{k}} \tag{4.6}$$

and, therefore,

$$\begin{aligned}\arg(1 + R(k)) &= \operatorname{Im}\ln(1 + R(k)) = -\operatorname{Im}\ln\frac{m(k) + i\sqrt{k}}{2i\sqrt{k}} \\ &= -\operatorname{Im}\ln\left(1 + \frac{m(k) - i\sqrt{k}}{2i\sqrt{k}}\right) \\ &= -\frac{1}{2}\operatorname{Im}\Omega(k) + \frac{1}{4k}\operatorname{Im}\Omega^2(k) + O\left(\frac{1}{k^{3/2}}\right),\end{aligned}$$

and hence, by Proposition 3.3, $\arg(1 + R(k)) \in L_1(a, \infty)$ with some $a > 0$. Since $|\arg(1 + R(k))| \le \pi$ for all k and $\left|\frac{z}{k - z}\right| \le 1, \operatorname{Re} z = 0$, by the Lebesgue dominated convergence theorem, we get (4.5). □

We do not know whether under condition (4.4) asymptotics (4.1) can be extended to the case when $k \to \infty$ along the real line, but at least the following holds.

THEOREM 4.3. *If $v(x) - c \in W_1^1(\mathbb{R}_+)$ for some c, then*

$$R(k) = \frac{v(0) - c}{4k} + o(1/k), \quad k \to \infty. \tag{4.7}$$

Proof. As before, set $c = 0$. From (4.6) and (2.10) we have

$$R(k) = \frac{1}{2i\sqrt{k}}\int_0^\infty e^{2i\sqrt{k}x} v(x)\, dx + o(1/k). \tag{4.8}$$

Let $v_\varepsilon(x)$ be a smooth function with a compact support containing $x = 0$ such that $v_\varepsilon \to v, \varepsilon \to 0$, in $W_1^1(\mathbb{R}_+)$, and $v_\varepsilon(0) = v(0)$. For the integral in (4.8) we then have

$$\begin{aligned}&\frac{1}{2i\sqrt{k}}\int_0^\infty e^{2i\sqrt{k}x} v_\varepsilon(x)\, dx + O\left(\|v - v_\varepsilon\|_1\right) = \frac{1}{4k}\left(v_\varepsilon(0) + \int_0^\infty e^{2i\sqrt{k}x} v_\varepsilon'(x)\, dx\right) \\ &\quad + O\left(\|v - v_\varepsilon\|_1\right) = \frac{1}{4k}v(0) + \frac{1}{4k}\int_0^\infty e^{2i\sqrt{k}x} v_\varepsilon'(x)\, dx + O\left(\|v - v_\varepsilon\|_{W_1^1}\right).\end{aligned}$$

Letting $\varepsilon \to 0$ yields

$$R(k) = \frac{v(0)}{4k} + \frac{1}{4k}\int_0^\infty e^{2i\sqrt{k}x} v_\varepsilon'(x)\,dx + o(1/k),$$

and by the Riemann–Lebesgue lemma, we arrive at (4.7). □

Remark 2. Under extra conditions $v \in L_1\left(\mathbb{R}_+, \left(1+x^2\right)dx\right)$ and the absence of resonances, (4.7) was also obtained in [18].

Remark 3. Summarizing the formulas of this section, we get a procedure of recovering potentials. Indeed, given point x_0 we cut off the potential to the left from x_0 and observe the plane waves reflected by the truncated potential. Knowing the reflection coefficient $R(x_0, k)$ for all energies k lets us now compute $v(x_0)$ by one of the formulas (4.1)–(4.3).

Remark 4. The setting of this section is a specific case of the abstract scheme of [22] (which also includes the three-dimensional case). In particular, it follows from [22] that $\frac{1}{\pi}\arg(1+R(k))$ appears as the spectral shift function of some self-adjoint operators.

5. Resonance states and potentials. As it follows from (2.13) the reflection coefficient $R(z), \operatorname{Im} z \geq 0$, is uniquely determined by $\left(1+k^2\right)^{-1}(\log|R(k)|dk + d\mu(k))$ on the real line and resonance states $\{k_n\}\left(k_n = z_n^2\right)$. One can easily observe from (4.1) that discarding a finite number of Blaschke factors $\frac{1+\overline{z}_j^2}{|1+z_j^2|}\frac{z-z_j}{z-\overline{z}_j}$ from $R(z)$ affects only the sign of $v(0)$. Actually, the same concerns some infinite Blaschke products. Let us say that a series of resonances $\{k_j\}$ does not contribute to computing the potential $|v|$ at $x = 0$ if

$$\left|\lim_{z\to i\infty} zR(z)\right| = \left|\lim_{z\to i\infty} zb^{-1}(z)\,R(z)\right|, \tag{5.1}$$

where $b(z)$ is the Blaschke product corresponding to $\{z_j\}$. The following proposition describes all series of resonances having property (5.1).

THEOREM 5.1. *Assume $v(x) \in L_1(0,\delta)$ for some $\delta > 0$. A series of resonances $\{k_j\}$ does not contribute to evaluating $|v(0)|$ by* (4.1) *if and only if $\{k_j\}$ satisfies*

$$\sum_j \frac{\operatorname{Im} z_j}{|z_j + i|} < \infty, \qquad \text{where } z_j^2 = k_j. \tag{5.2}$$

Proof. Assume that $\{z_j\}$ is subject to (5.2). It follows from the upper half-plane version of the Frostman theorem [4] that

$$b(z) \equiv \prod_j \frac{1+\overline{z}_j^2}{|1+z_j^2|}\frac{z-z_j}{z-\overline{z_j}}$$

has a limit as $z \to i\infty$ along the imaginary axis and $|\lim_{\tau\to\infty} b(i\tau)| = 1$ that immediately implies (5.1).

Now suppose $\{k_j\}$ do not contribute to the limit (4.1). It follows from (5.1) that $\lim_{\tau\to\infty} b(i\tau)$ exists and is equal by modulus to 1. Applying the same Frostman theorem to $\{z_j\}$ yields (5.2). □

In complex analysis, condition (5.2) is referred to as the Frostman condition at infinity [4]. Note that if $\{z_j\}$ accumulates at a finite distance, then the Blaschke condition (2.14) for $\{z_j\}$ is equivalent to (5.2). Since $\left(\operatorname{Im}\sqrt{z_j}\right)^{-1}$ defines the lifetime

of k_j, condition (5.2) also means that series of resonances subject to (5.2) are in a certain sense long living.

THEOREM 5.2. *If a potential $v(x)$ is such that $v(0) \neq 0$ and the spectrum of H is purely point, then*

$$\sum_j \frac{\operatorname{Im} z_j}{|z_j + i|} = \infty.$$

Proof. Since $\sigma(H) = \sigma_p(H)$, the reflection coefficient $|R(k)| = 1$ for almost all real k. Moreover, it follows from (2.12) that $R(z)$ is analytic on $\mathbb{R} \setminus \sigma(H)$ and at any point $k_0 \in \sigma_p(H)$ we have $\lim_{z \to k_0} R(z) = -1, z \in \mathbb{C}_+$. Therefore, by [7] the measure μ in (2.13) is 0, and hence $R(z)$ is a product of $\exp iaz$ and a Blaschke product. Since $\lim z \exp iaz = 0, z \to i\infty$, contradicts the condition $v(0) \neq 0$ we conclude that $a = 0$, and hence $R(z) = B(z)$, where $B(z)$ is a Blaschke product. However, it follows from (4.1) that $\lim_{z \to i\infty} R(z) = 0$, which implies that $B(z)$ must be an infinite Blaschke product with the property $\lim_{\tau \to \infty} B(i\tau) = 0$. By the Frostman theorem [4] this means that (5.2) fails and

$$\sum_j \frac{\operatorname{Im} z_j}{|z_j + i|} = \infty.$$

This proves the theorem. □

6. The trace formula. In this section we discuss the situation when we no longer cut off the left side of a potential and substitute it for 0. The reflection coefficient can be defined even in this setting, but we prefer to follow [8, 9, 10, 11, 12, 13, 14, 15], and as scattering data we now consider Krein's spectral shift function $\xi(x_0, k)$ of the pair (H_{x_0}, H). The main objective of this section is to extend some of the results of Gesztesy et al. [8, 11], Gesztesy [9], Gesztesy and Holden [10, 15], Gesztesy, Holden, and Simon [12], and Gesztesy and Simon [13, 14]. In the literature, formulas of type (1.1) are related to the so-called trace approach to the inverse scattering problem. Under different (but rather restrictive) conditions and various forms, they have been previously studied by many authors. A good account on the literature can be found, e.g., in [12].

THEOREM 6.1 (trace formula). *Suppose $v(x) \in L_1(x_0 - \delta, x_0 + \delta)$ for some $\delta > 0$. If x_0 is both a left and right Lebesgue point, then*

$$\frac{v(x_0 - 0) + v(x_0 + 0)}{2} = \lim_{z \to i\infty} \int_{\mathbb{R}} \left(\frac{z}{k - z} \right)^2 \{\chi(k) - 2\xi(x_0, k)\} \, dk, \tag{6.1}$$

where $\chi(k)$ is the characteristic function of $\mathbb{R}_+$.

Proof. Without loss of generality we set $x_0 = 0$. It follows from Proposition 2.3 that for $z \to i\infty$

(6.2)

$$\frac{m_+(z) - m_-(z)}{2i\sqrt{z}} = 1 - \frac{1}{2i\sqrt{z}} \left(\int_0^\delta e^{2i\sqrt{z}x} v(x) \, dx + \int_{-\delta}^0 e^{-2i\sqrt{z}x} v(x) \, dx \right) + o\left(1/\sqrt{z}\right).$$

Functions $m_+(z) - m_-(z)$ and $2i\sqrt{z}$ are clearly Herglotz, and by (2.2) we have

$$m_+(z) - m_-(z) = \exp\left\{ a + \int_{\mathbb{R}} \left(\frac{1}{k - z} - \frac{k}{1 + k^2} \right) \alpha(k) dk \right\}, \tag{6.3}$$

$$2i\sqrt{z} = \exp\left\{ \ln 2 + \int_{\mathbb{R}} \left(\frac{1}{k - z} - \frac{k}{1 + k^2} \right) \beta(k) dk \right\},$$

where a is an inessential constant and

$$\alpha(k) = \frac{1}{\pi} \lim_{\varepsilon \to o+} \arg \{m_+ (k + i\varepsilon) - m_- (k + i\varepsilon)\}, \tag{6.4}$$
$$\beta(k) = 1, k < 0; \quad \beta(k) = 1/2, k > 0.$$

Plugging (6.3) into (6.2) and passing to the logarithm, we get

$$a + \int_{\mathbb{R}} \left(\frac{1}{k - z} - \frac{k}{1 + k^2} \right) (\alpha(k) - \beta(k))\, dk \tag{6.5}$$
$$= \ln \left\{ 1 - \frac{1}{2i\sqrt{z}} \left(\int_0^\delta e^{2i\sqrt{z}x} v(x)\, dx + \int_{-\delta}^0 e^{-2i\sqrt{z}x} v(x)\, dx \right) + o\left(1/\sqrt{z}\right) \right\}$$
$$= -\frac{1}{2i\sqrt{z}} \left(\int_0^\delta e^{2i\sqrt{z}x} v(x)\, dx + \int_{-\delta}^0 e^{-2i\sqrt{z}x} v(x)\, dx \right) + o\left(1/\sqrt{z}\right).$$

Now relate $\alpha(k)$ to the Krein spectral shift function $\xi(k)$ of the pair (H_0, H). Since (see, e.g., [13])

$$\xi(k) = \frac{1}{\pi} \arg \{m_- (k + i0) - m_+ (k + i0)\}^{-1} \tag{6.6}$$

and is chosen to be $0 \le \xi(k) \le 1$, one immediately has $\alpha(k) + \xi(k) = 1$, and hence

$$\alpha(k) - \beta(k) = -\xi(k) + 1/2\chi(k).$$

Equation (6.5) now reads

$$a + \int_{\mathbb{R}} \left(\frac{1}{k - z} - \frac{k}{1 + k^2} \right) \left(\frac{\chi(k)}{2} - \xi(k) \right) dk \tag{6.7}$$
$$= -\frac{1}{2i\sqrt{z}} \int_0^\delta e^{2i\sqrt{z}x} \{v(x) + v(-x)\}\, dx + o\left(1/\sqrt{z}\right).$$

Applying Lemma 3.2 to the right-hand side of (6.7) one gets

$$a + \int_{\mathbb{R}} \left(\frac{1}{k - z} - \frac{k}{1 + k^2} \right) \left(\frac{\chi(k)}{2} - \xi(k) \right) dk = -\frac{v(+0) + v(-0)}{4z} + o\left(\frac{1}{z}\right),$$

which, due to Lemma 3.1, yields (6.1) for $x_0 = 0$. □

Remark 5. Under the additional hypothesis that $v(x) \ge c > -\infty$ and x_0 is a Lebesgue point of $v(x)$, formula (6.1) turns into

$$v(x_0) = E + \lim_{z \to i\infty} \int_E^\infty \left(\frac{z}{k - z} \right)^2 \{1 - 2\xi(x_0, k)\}\, dk, \tag{6.8}$$

which is (1.1). The original proof of (6.8) employs a different strategy based upon the Krein trace formula (2.15) for

$$\operatorname{tr}\{\exp(-tH_{x_0}) - \exp(-tH)\} \tag{6.9}$$

and the Feynman–Kac formula to derive the asymptotic expansion of (6.9) as $t \to +0$. In this way, the abovementioned authors first obtain

$$v(x_0) = E + \lim_{t \to +0} \int_E^\infty e^{-kt} \{1 - 2\xi(x_0, k)\}\, dk, \tag{6.10}$$

which is valid only if v is essentially bounded from below. Formula (6.8) then follows from (6.10) and can be clearly interpreted as a resolvent regularization of the diverging trace $\mathrm{tr}(H_{x_0}-H)$. Now it is clear why (6.1) is called a trace formula. Note that under our conditions formula (6.10) no longer holds.

Let us also note that a trace formula close to (6.1) was recently obtained in [16] as a corollary of some general results on the operator-valued version of (6.8), but the requirement of the setting imposes $v\in L_\infty(\mathbb{R})\cap C[x_0-\delta,x_0+\delta]$.

Theorem 6.1 can also be stated for other than Dirichlet boundary conditions. Let $H^h_{x_0}=H^{h,+}_{x_0}\oplus H^{h,-}_{x_0}$, where $h\in\mathbb{R}$ and

$$\begin{aligned}H^{h,\pm}_{x_0}u&=lu, u\in \mathrm{Dom}\left(H^{h,\pm}_{x_0}\right)\\ &=\left\{f\in L_2(x_0,\pm\infty): lf\in L_2(x_0,\pm\infty), f'(x_0\pm 0)+hf(x_0\pm 0)=0\right\}.\end{aligned}$$

THEOREM 6.2. *Suppose a real-valued $v(x)\in L_1(x_0-\delta,x_0+\delta)$ and x_0 is both a left and right Lebesgue point. If $\xi_h(x_0,k)$ is the spectral shift function of the pair $(H,H^h_{x_0})$, then*

$$\frac{v(x_0-0)+v(x_0+0)}{2}=2h^2+\lim_{z\to i\infty}\int_{\mathbb{R}}\left(\frac{z}{k-z}\right)^2\{\chi(k)+2\xi_h(x_0,k)\}dk. \tag{6.11}$$

Proof. As in Theorem 6.1 one can treat only the case $x_0=0$. Let $m_{h,\pm}(z)=m_{h,\pm}(z)$ be Weyl's m-functions corresponding to $\mathbb{R}_\pm$ and associated with the boundary condition $u'(\pm 0)+hu(\pm 0)=0$. By (2.11) we have

$$m_{h,+}-m_{h,-}=\frac{\left(1+h^2\right)(m_+-m_-)}{(m_++h)(m_-+h)}. \tag{6.12}$$

Rewrite (6.12) as follows:

$$\frac{-i\sqrt{z}}{2(1+h^2)}(m_{h,+}-m_{h,-})=\frac{\frac{m_+-m_-}{2i\sqrt{z}}}{\frac{m_++h}{i\sqrt{z}}\cdot\frac{m_-+h}{-i\sqrt{z}}}.$$

The function $m_{h,+}-m_{h,-}$ is Herglotz and applying the very same arguments as in the proof of Theorem 6.1 yields

$$\begin{aligned}&a(h)+\int_{\mathbb{R}}\left(\frac{1}{k-z}-\frac{k}{1+k^2}\right)(\alpha_h(k)+\beta(k))\,dk\\ &=\ln\frac{m_+-m_-}{2i\sqrt{z}}-\ln\left(\frac{m_++h}{i\sqrt{z}}\cdot\frac{m_-+h}{-i\sqrt{z}}\right)\end{aligned} \tag{6.13}$$

with some inessential complex constant $a(h)$. By (2.10) and (2.11) we have

$$\frac{m_\pm+h}{\pm i\sqrt{z}}=1\pm\frac{h}{i\sqrt{z}}\mp\frac{1}{i\sqrt{z}}\int_0^{\pm\delta}e^{\pm 2i\sqrt{z}x}v(x)\,dx+o\left(1/\sqrt{z}\right),\quad z\to i\infty. \tag{6.14}$$

In analogy to (6.8) one gets $\alpha_h(k)+\beta(k)=-\left(\xi_h(k)+(1/2)\chi(k)\right)$. It follows from (6.13), (6.14) that

$$\begin{aligned}&a(h)+\int_{\mathbb{R}}\left(\frac{1}{k-z}-\frac{k}{1+k^2}\right)\left(\xi_h(k)\ +\frac{\chi(k)}{2}\right)dk\\ &\qquad=-\frac{v(-0)+v(+0)-8h^2}{4z}+o(1/z),\quad z\to i\infty,\end{aligned}$$

and Lemma 3.2 immediately applies (6.11). □

Note that Theorem 6.2 extends some of the results of [8, 9, 10, 11], where v is assumed to be bounded from below.

7. Absolute summability of the trace formula. Now discuss conditions when we can remove the resolvent regularization in (6.1). As the following theorem shows, it can be done for certain potential with different spatial asymptotics at $\pm\infty$.

THEOREM 7.1. *Suppose $v(x)$ is subject to*

$$\exists v_\pm \in \mathbb{R} : v(x) - v_\pm \in L_1(\mathbb{R}_\pm), \quad \int_{\mathbb{R}_\pm} (v(x+x_0) - v_\pm) \cos kx \cdot dx \in L_1(\mathbb{R}). \tag{7.1}$$

If $\xi(x_0, k)$ is the spectral shift function of the pair (H_{x_0}, H), then

$$\frac{v(x_0 - 0) + v(x_0 + 0)}{2} = E + \int_E^\infty (1 - 2\xi(x_0, k))\, dk, \tag{7.2}$$

where $E = \inf \sigma(H)$.

Proof. As before, it is enough to treat the case $x_0 = 0$. Note first that under hypothesis (7.1) the spectrum of H is bounded from below by $E = \inf \sigma(H)$, and formula (6.1) then reads

$$\frac{v(-0) + v(+0)}{2} = E + \lim_{z \to i\infty} \int_E^\infty \left(\frac{z}{k - z}\right)^2 \{1 - 2\xi(k)\}\, dk.$$

It follows from (6.4) and (6.6) that

$$1 - 2\xi(k) = \frac{1}{\pi} \operatorname{Im} \ln \frac{m_+(k) - m_-(k)}{i\sqrt{k}}. \tag{7.3}$$

Let us evaluate (7.3). Setting $k_\pm = \sqrt{k - v_\pm}$ we have

$$\begin{aligned} \frac{m_+(k) - m_-(k)}{i\sqrt{k}} &= \frac{m_+(k) - ik_+}{i\sqrt{k}} + \frac{m_-(k) + ik_-}{-i\sqrt{k}} + \frac{k_+ + k_-}{\sqrt{k}} \\ &= \frac{k_+ + k_-}{\sqrt{k}} \left\{1 + \frac{\sqrt{k}}{k_+ + k_-} (\Omega_+(k) + \Omega_-(k))\right\}, \end{aligned}$$

where $\Omega_\pm(k)$ as in Proposition 3.3. It follows then from (7.3) that

$$1 - 2\xi(k) = \frac{1}{\pi} \operatorname{Im} \ln \left\{1 + \frac{\sqrt{k}}{k_+ + k_-} (\Omega_+(k) + \Omega_-(k))\right\}$$

for $k \geq k_0 = \max\{v_+, v_-\}$. In virtue of Proposition 3.3 we have $\Omega_+(k) + \Omega_-(k) \in L_1(a, \infty)$ with some $a > k_0$. On the other hand, from (6.4) and (6.6) one has $|1 - 2\xi(k)| \leq 1$ for all $k \in \mathbb{R}$. However, $\left|\frac{z}{k-z}\right| \leq 1, \operatorname{Re} z = 0$, and applying Lebesgue's dominated convergence theorem then completes the proof. ∎

Theorem 7.1 appears to be first obtained in [6] in a different form under pretty strong requirements on potentials v. Another proof was recently presented in [12] assuming v, v' are locally absolute continuous, $v - v_\pm \in L_1(\mathbb{R}_\pm)$, for some real constants $v_\pm$, and $v', v'' \in L_1(\mathbb{R})$. Let us note that our arguments allow one to conduct the proof under sharp conditions on v. Indeed, the only thing we rely on is Proposition 3.3, where the conditions on $v(x)$ are clearly sharp (and might even be necessary/sufficient). What (7.1) actually says is that besides $v - v_\pm \in L_1(\mathbb{R}_\pm)$ potentials v must be slightly better than locally continuous on $(-\infty, x_0)$ and (x_0, ∞), e.g., Dini continuous (see, e.g., [27]). Also note that we do allow for one point of jump discontinuity at x_0. Indeed, conditions of Theorem 7.1 force v to be continuous on $(-\infty, x_0)$ and (x_0, ∞), but the condition $v(x_0 + 0) = v(x_0 - 0)$ need not hold.

REFERENCES

[1] N. Aronszajn and W.F. Donoghue, *On exponential representation of analytic functions in the upper half-plane with positive imaginary part*, J. Anal. Math. 5 (1957), pp. 321–388.

[2] F. Atkinson, *On the location of the Weyl circles*, Proc. Roy. Soc. Edinburgh Sect. A, 88 (1981), pp. 345–356.

[3] M.Sh. Birman and D.R. Yafaev, *The spectral shift function. The work of M.G. Krein and its further development*, St. Petersburg Math. J., 4 (1993), pp. 833–870.

[4] E.F. Collingwood and A.J. Lohwater, *The Theory of Cluster Sets*, Cambridge University Press, London, 1966.

[5] W. Craig, *The trace formula for Schrödinger operators on the line*, Comm. Math. Phys, 126 (1989), pp. 379–407.

[6] P. Deift and E. Trubowitz, *Inverse scattering on the line*, Comm. Pure Appl. Math., 32 (1979), pp. 121–251.

[7] J.B. Garnett, *Bounded Analytic Functions*, Academic Press, New York, 1981.

[8] F. Gesztesy, H. Holden, B. Simon, and Z. Zhao, *Trace formulae and inverse spectral theory for Schrödinger operators*, Bull. Amer. Math. Soc. (N.S.), 29 (1993), pp. 250–255.

[9] F. Gesztesy, *New trace formulas for Schrödinger operators*, in Evolution Equations, G. Ferreyra, G. Goldstein, and F. Neubrander, eds., Marcel Dekker, New York, 1995, pp. 201–221.

[10] F. Gesztesy and H. Holden, *On new trace formulae for Schrödinger operators*, Acta Appl. Math., 39 (1995), pp. 315–333.

[11] F. Gesztesy, H. Holden, B. Simon, and Z. Zhao, *Higher order trace relations for Schrödinger operators*, Rev. Math. Phys., (1995), pp. 893–922.

[12] F. Gesztesy, H. Holden, and B. Simon, *Absolute summability of the trace relation for certain Schrödinger operators*, Comm. Math. Phys., 168 (1995), pp. 137–161.

[13] F. Gesztesy and B. Simon, *The ξ function*, Acta Math., 176 (1996), pp. 49–71.

[14] F. Gesztesy and B. Simon, *Uniqueness theorems in inverse spectral theory for one-dimensional Schrödinger operators*, Trans. Amer. Math. Soc., 348 (1996), pp. 349–373.

[15] F. Gesztesy and H. Holden, *On trace formulas for Schrödinger-type operators*, in Multiparticle Quantum Scattering with Applications to Nuclear, Atomic and Molecular Physics, D.G. Truhlar and B. Simon, eds., Springer, New York, 1997, pp. 121–145.

[16] F. Gesztesy and K. Makarov, *Some applications of the spectral shift operator*, in Operator Theory and its Applications, A.G. Ramm, P.N. Shivakumar, and A.V. Strauss, eds., Fields Inst. Commun. 25, AMS, Providence RI, 2000, pp. 267–292.

[17] B.J. Harris, *The asymptotic form of the Titchmarsh–Weyl m-function associated with a second order differential equation with locally integrable coefficients*, Proc. Roy. Soc. Edinburgh Sect. A, 102 (1986), pp. 242–251.

[18] M.V. Klibanov and P.E. Sacks, *Phaseless inverse scattering and the phase problem in optics*, J. Math. Phys., 33 (1992), pp. 3813–3821.

[19] B.M. Levitan and I.S. Sargsjan, *Introduction to Spectral Theory*, AMS, Providence, RI, 1975.

[20] B.S. Pavlov, *On the one-dimensional scattering of plane waves on an arbitrary potential*, Theoret. and Math. Phys., 16 (1973), pp. 706–713.

[21] A.V. Rybkin, *KdV invariants and Herglotz functions*, Differential Integral Equations, 14 (2001), pp. 493–512.

[22] A.V. Rybkin, *On A-integrability of the spectral shift function of unitary operators arising in the Lax-Phillips scattering theory*, Duke Math. J., 83 (1996), pp. 683–699.

[23] A.V. Rybkin, *A Weyl m-function Approach to the Trace Formulas for Various Schrödinger Operators*, in preparation.

[24] B. Simon, *A new approach to inverse spectral theory,* I. *Fundamental formalism*, Ann. of Math. (2), 150 (1999), pp. 1029–1057.

[25] E.C. Titchmarsh, *On Eigenfunction Expansions Associated with Second-Order Differential Equations*, Oxford University Press, New York, 1950.

[26] S. Venakides, *The infinite period limit of the inverse formalism for periodic potentials*, Comm. Pure Appl. Math., 41 (1988), pp. 3–17.

[27] A. Zygmund, *Trigonometric Series*, Vol. 2, Cambridge University Press, New York, 1959.

SIAM J. MATH. ANAL.
Vol. 32, No. 6, pp. 1265–1271

ASYMPTOTICALLY-FREE SOLUTIONS FOR THE SHORT-RANGE NONLINEAR SCHRÖDINGER EQUATION*

KENJI NAKANISHI†

Abstract. We consider the nonlinear Schrödinger equation (NLS) with a power nonlinearity $|u|^{p-1}u$, where $1+2/n<p$. We show that for any free solution in L^2 or H^1 there exists a solution of NLS which approaches the free solution in the same space as t tends to infinity. We also show that the wave operators exist in $\Sigma^s=H^s\cap\mathcal{F}H^s$ for the critical power $p=1+4/(n+2s)$.

Key words. nonlinear Schrödinger equation, scattering

AMS subject classifications. 35Q55, 35P25

PII. S0036141000369083

1. Introduction and main result. In this note, we study asymptotic behavior of the solutions for the nonlinear Schrödinger equation (NLS):

$$i\dot{u}-\Delta u+\lambda|u|^{p-1}u=0, \tag{1.1}$$

where $u=u(t,x):\mathbb{R}^{1+n}\to\mathbb{C}$, $\dot{u}=\partial u/\partial t$, $n\in\mathbb{N}$, $\lambda\in\mathbb{R}$, and $p>1+2/n$. There is a large amount of literature on the scattering theory for the equation above. It is well known that we need $p>1+2/n$ to have the wave operators. In this short-range case, there are many results on the existence of the wave operators and the scattering operator. However, the available results on the scattering in the energy space are restricted to the case where $p\geq 1+4/n$, and those results concerned with the case $1+2/n<p<1+4/n$ rely totally on decay assumptions for initial data at the spatial infinity. Such decay conditions are not compatible with the Hamiltonian (conservative) structure of the equation and the free propagator is no longer unitary on the function spaces with the corresponding weights. When we consider the scattering theory in such spaces, the space-decay provides a priori time-decay to the solutions and enables us to use perturbative arguments at the time infinity for solutions with sufficient decay in time. Thus, our understanding so far is very limited concerning the asymptotic behavior of the solutions along the conservative structure or the unitary-free evolution if $p<1+4/n$. Here we present a first step for this question; we show that for any free solution in L^2 or H^1 there exists a solution of NLS in the same space which is asymptotic to the free solution. The uniqueness of such a solution would give us well-defined wave operators in those spaces, but we have no idea if the uniqueness holds or not. It is generally very difficult to get uniqueness only from the information that the solutions approach to a fixed free solution (as in (1.2)).

It is an advantage of our result that we do not need any restriction for higher dimensions as in the case of the scattering in weighted spaces, though we have not succeeded in treating the low dimensional case, which is easier for the weighted scattering. Our main result follows.

*Received by the editors March 13, 2000; accepted for publication (in revised form) October 26, 2000; published electronically March 15, 2001. This author's research was supported by Research Fellowships of the Japan Society for the Promotion of Science for Young Scientists.
http://www.siam.org/journals/sima/32-6/36908.html
†Department of Mathematics, Kobe University, Rokko, Kobe 657-8501, Japan (kenji@math.kobe-u.ac.jp).

THEOREM 1.1. *Let $n \geq 3$ and $1 + 2/n < p < 1 + 4/n$. Define $X = L^2(\mathbb{R}^n)$ or $H^1(\mathbb{R}^n)$. Then for any solution $v \in C(\mathbb{R}; X)$ of the free Schrödinger equation, there exists a solution $u \in C(\mathbb{R}; X)$ of NLS satisfying*

$$\lim_{t\to\infty} \|u(t) - v(t)\|_X = 0. \tag{1.2}$$

Let us compare this with the known results. As for the converse correspondence, i.e., from u to v, it is known that $U(-t)u(t)$ converges weakly in H^1 for any finite energy solution of NLS, where $U(t) = e^{-it\Delta}$ [15]. A similar result for the nonlinear Klein–Gordon equation was proved in [11]. Notice that there is significant difference between the weak and the strong convergence in this problem, since the weak topology cannot even distinguish the standing waves from the 0 solution. However, if we assume additionally that $xu(0) \in L^2$ and $\lambda \geq 0$, then it is known that $U(-t)u(t)$ converges strongly in L^2 [18].

If $1 + 4/n < p < 1 + 4/(n-2)$, $\lambda \geq 0$, and $X = H^1$, then the solution u above is uniquely determined so that we have the wave operators well defined on the whole energy space for any $n \in \mathbb{N}$. Moreover, we know the asymptotic completeness, i.e., $v(0) \mapsto u(0)$ is a bijection on the energy space H^1 [10, 8, 12]. A similar result is given in [3] for radial data in the case where $p = 1 + 4/(n-2)$ and $n \geq 3$.

When $p = 1 + 4/n$, we can define the wave operators on $X = H^1$ if we restrict the nonlinear solutions to those possessing a certain global space-time integrability. It is an open problem whether every finite energy solution has that property when $\lambda \geq 0$, which would also imply the asymptotic completeness; the invertibility of the wave operators is known only for small L^2 data or in some weighted spaces [7, 8, 2].

For $p < 1 + 4/n$, most results are obtained in weighted spaces. For $0 < s < 2$, the wave operators are well defined on $\Sigma^s = H^s \cap \mathcal{F}H^s$ if $p > \max(1+2/n, 1+4/(n+2s), s)$, where $\mathcal{F}$ denotes the Fourier transform [4, 6]. Here we remark that $p = 1+4/(n+2s)$ is allowed if it is larger than $1+2/n$ and s. We will show this fact in the next section. If $p \leq 1 + 2/n$, then the wave operators cannot exist [1, 9, 15], but the modified wave operators are given in some cases [13, 5]. If $p \geq 1 + 8/(\sqrt{(n+2)^2 + 8n} + n - 2)$ and $\lambda \geq 0$, then we have the asymptotic completeness in Σ^1 [7, 16, 4].

The proof of the above theorem relies on the compactness argument. To have compatibility of the convergence and the nonlinearity, we use the local smoothing effect of the Schrödinger equation. The strong convergence follows from the weak convergence and the conservation laws. The condition $n \geq 3$ is required so that the decay order $t^{-n/2}$ of the free propagator is integrable for $t \to \infty$. This fact is used only for the weak continuity in the compactness argument. However, it might be too optimistic to regard that condition as solely technical, because the above result is indeed quite delicate, which can be seen by the following observation. Let u and v be as in the above theorem. Then, by the equation NLS we have

$$\langle iu, v\rangle(T) - \langle iu, v\rangle(S) = -\int_S^T \langle \lambda|u|^{p-1}u, v\rangle(t)dt, \tag{1.3}$$

where $\langle u, v\rangle := \Re \int u\bar{v}dx$. Since u approaches to v as $t \to \infty$, we might try to replace u by v in the right-hand side. However, it is easy to construct a free solution v with finite energy satisfying

$$\int_S^\infty \int |v|^{p+1}dxdt = \infty \tag{1.4}$$

for any $S > 0$ if $p < 1+4/n$. Nevertheless, the left-hand side of (1.3) converges to 0 as $S, T \to \infty$. Thus, u does not approximate v in such a strong sense but is even better than v itself in the meaning above. Notice that the integral is finite if $p \geq 1 + 4/n$ and also for the nonlinear solution u if $p > 1 + 4/n$, due to the Strichartz estimate and the scattering result. It is also an interesting open problem if the integral is finite for the finite energy solution u obtained in the above theorem.

Another way to observe the difficulty of this problem is to use the pseudoconformal transform. Our problem is converted by the transform into the following Cauchy problem:

$$\begin{aligned} i\dot{u} - \Delta u + \lambda t^{-\nu}|u|^{p-1}u = 0, \\ u(0) = \overline{\mathcal{F}v(0)}, \end{aligned} \tag{1.5}$$

where $\nu = n(1+4/n-p)/2$. Then the scaling argument tells us that this problem with $X = L^2$ is in the so-called "super-critical" case, as is the nonlinear wave equation on $\mathbb{R}^{1+3}$ with a power greater than 5 and the Navier–Stokes equation on $\mathbb{R}^{1+3}$. Namely, if we expand a solution by the scaling which leaves the equation invariant, then the initial norm (in L^2 for our problem) increases. This means that the Cauchy problem does not become easier at all even if we restrict the argument to a smaller time interval or small data. However, our result is better than those about the other examples above, since we have the strong continuity in time and the conservation laws for the solutions that we obtain.

Proof of Theorem 1.1. By the Strichartz estimate, for any $T > 0$ we have the unique solution w to

$$\begin{aligned} i\dot{w} - \Delta w + \lambda|w|^{p-1}w = 0, \\ w(T) = v(T), \end{aligned} \tag{1.6}$$

satisfying $w \in C(\mathbb{R}; X) \cap Y$, where Y denotes the Banach space endowed with the following norm:

$$\|w\|_Y := \sup_{\tau\in\mathbb{R}} \|w\|_{L^{2+4/n}((\tau,\tau+1)\times\mathbb{R}^n)}. \tag{1.7}$$

See [17]. In the following, the arguments about energy are concerned only with the case $X = H^1$. w satisfies the conservation laws for the charge and the energy:

$$\|w(t)\|_{L^2} = \|v(0)\|_{L^2} < \infty, \tag{1.8}$$

$$\begin{aligned} E(w;t) &:= \int |\nabla w(t)|^2 + \frac{2\lambda}{p+1}|w(t)|^{p+1}dx \\ &= \int |\nabla v(T)|^2 + \frac{2\lambda}{p+1}|v(T)|^{p+1}dx. \end{aligned} \tag{1.9}$$

Thus, the charge does not depend on T and

$$E(w;t) = E(w;0) \to \int |\nabla v(0)|^2 dx =: E(v) \tag{1.10}$$

as $T \to \infty$ because we have

$$\|v(T)\|_{L^{p+1}} \to 0 \tag{1.11}$$

for any free solution v with finite energy since $2 < p+1 < 2n/(n-2)$. By the Sobolev inequality and the interpolation, we have

$$\int |w(t)|^{p+1}dx \le C\|\nabla w(t)\|_{L^2}^{\alpha}\|w(t)\|_{L^2}^{p+1-\alpha}, \tag{1.12}$$

where $\alpha < 2$ since $p < 1 + 4/n$. Thus we have a uniform bound of $w(t)$ in X. Now we take the weak limit $T \to \infty$. Let $A \subset C_0^\infty(\mathbb{R}^n)$ be an enumerable set which is dense in L^2. We denote $U(t) = e^{-it\Delta}$. Then, for any $\psi \in A$, the L^2 coupling $(U(-t)w(t), \psi)$ is a function of t bounded uniformly for t and T. We also have the uniform equicontinuity up to $t \to \infty$ as follows:

$$\begin{aligned} |(U(-t)w(t),\psi) - (U(-s)w(s),\psi)| &= \left|\int_s^t (U(-\sigma)\lambda|w|^{p-1}w(\sigma),\psi)d\sigma\right| \\ &\le C\int_s^t \sigma^{-1-\varepsilon}\|w(\sigma)\|_{L^q}^p\|\psi\|_{L^r}d\sigma \\ &\le C\|v(0)\|_{L^2}^{(1-\theta)p}\|w\|_Y^{\theta p}\|\psi\|_{L^r}\min(s^{-\varepsilon}, |t-s|^{1-p/q}) \\ &\le C(\|v(0)\|_{L^2},\psi)\min(s^{-\varepsilon}, |t-s|^{1-p/q}), \end{aligned} \tag{1.13}$$

where $1 < s < t$, $1/r = p/q = 1/2 + (1+\varepsilon)/n$, and $\varepsilon > 0$ is a constant chosen such that $1/q = (1-\theta)/2 + \theta/(2+4/n)$ with $0 < \theta < 1$ and $p/q < 1$, which is possible since $1 + 4/n > p > 1 + 2/n$ and $n \ge 3$. Thus, $\{(U(-t)w(t),\psi)\}_{T>0}$ is uniformly bounded and equicontinuous on $t \in [-\infty, \infty]$. Now we can apply the Ascoli–Arzela theorem and by a further diagonal argument, we obtain a subsequence w such that $(U(-t)w(t),\psi)$ converges uniformly on $t \in [-\infty,\infty]$ as $T \to \infty$ for any $\psi \in A$. By the X boundedness of w and the denseness of A in the dual space X', w converges weakly in X. Denote the limit by u. The limit function $U(-t)u(t)$ is weakly continuous on $[-\infty,\infty]$ and u satisfies $U(-t)u(t) \to v(0)$ as $t \to \infty$ weakly in X. To show that u solves (1.1), we need that $f(w)$ converges to the appropriate limit $f(u)$, for which purpose the weak topology is not efficient in general. When $X = H^1$, we know the standard argument via the compactness of the Sobolev embedding to ensure such compatibility. Even if $X = L^2$, we can argue similarly in the space-time by the local smoothing effect. In fact, we have the following.

LEMMA 1.2. *Let $u_\nu \in C(\mathbb{R}; L^2) \cap Y$ be a sequence of solutions to* (1.1) *with bounded L^2 norms (Y is defined in* (1.7)*). Then, u_ν converges along some subsequence in $L^q_{loc}(\mathbb{R}^{1+n})$ for any $q < 2 + 4/n$.*

We postpone the proof of this lemma and continue the proof of the theorem. Since $p < 2 + 4/n$, the above lemma implies that $f(w)$ converges to $f(u)$ in $L^1_{loc}(\mathbb{R}^{1+n})$, so that the limit function u is a weak solution of (1.1) in Y. Since we have the uniqueness of such a solution to the Cauchy problem for NLS, this solution u satisfies the conservation laws. Because it is a weak limit as $T \to \infty$ in L^2 at each t, we have

$$\|U(-t)u(t)\|_{L^2} = \|u(t)\|_{L^2} =: C_\infty \le \|v(0)\|_{L^2}, \tag{1.14}$$

but we know that its weak limit as $t \to \infty$ is exactly $v(0)$, which implies that $C_\infty = \|v(0)\|_{L^2}$ and all the convergences in L^2 must be strong. In the case of $X = H^1$, from the L^2 convergence and the H^1 boundedness, we have the L^{p+1} convergence as $T \to \infty$. Then, from (1.9), (1.10), and (1.11) we have

$$E(u;t) =: E_\infty \le \int |\nabla v(0)|^2 dx = E(v). \tag{1.15}$$

Since $U(-t)u(t) \to v(0)$ weakly in H^1, we have

$$E(v) \le \liminf_{t\to\infty} \|\nabla U(-t)u(t)\|_{L^2}^2 \le E_\infty. \tag{1.16}$$

Therefore, $E_\infty = E(v)$ and all the convergences must be strong in H^1. Thus we obtain the desired solution $u \in C(\mathbb{R}; X)$ satisfying $U(-t)u(t) \to v(0)$ strongly in X. □

Remark 1.3. The subsequence of w chosen above converges in $C(\mathbb{R}; X)$. If we have the uniqueness of the limit, then a similar argument will also yield the continuity of the wave operators.

Proof of Lemma 1.2. We use the local smoothing effect of the Schrödinger equation [14]

$$\|\chi U(t)\varphi\|_{L_t^2 H_x^{1/2}} \le C\|\varphi\|_{L^2}, \tag{1.17}$$

where χ is an arbitrary function in $C_0^\infty(\mathbb{R}^{1+n})$ and C is a positive constant dependent only on χ. Interpolating this with the Strichartz estimate, we obtain

$$\left\|\chi \int_0^t U(t-s)F(s)ds\right\|_{L_t^\alpha H_x^s} \le C\|F\|_{L_t^\beta L_x^\gamma}, \tag{1.18}$$

for $1 \le \beta < 2$, $n/\gamma - n/2 = 2 - 2/\beta$ and $1/\alpha > 1/\beta - 1/2 > s$, where C is a positive constant dependent only on these exponents and χ. Since the L^2 bound implies the boundedness in Y, we can use the above estimates to deduce that χu_ν are also bounded in $L_t^2 H_x^s$ for some $s > 0$. By the equation NLS, $\chi \dot{u}_\nu$ are bounded in $L_t^2 H_x^{-2}$. Hence, by the interpolation, there exists some $\delta > 0$ such that χu_ν are bounded in $H^\delta(\mathbb{R}^{1+n})$. By the compactness of the Sobolev embedding, u_ν converges along some subsequence in $L^2_{loc}(\mathbb{R}^{1+n})$. Then, by the $L^{2+4/n}$ boundedness, it converges also in $L^q_{loc}(\mathbb{R}^{1+n})$ for any $q < 2 + 4/n$. □

2. Wave operators for the critical power on Σ^s. In this section we show that the wave operators are well defined on $\Sigma^s = H^s \cap \mathcal{F}H^s$ for $p = 1 + 4/(n + 2s)$ if $p > 1 + 2/n$ and $p > s > 0$.

Following [16, 4], the construction of the wave operators is converted into the Cauchy problem (1.5) with $\mathcal{F}v(0) \in \Sigma^s$. Then, the argument to solve the Cauchy problem is quite standard with the fixed point theorem and the Strichartz estimate. The main point is to derive a closed estimate for the inhomogeneous term in the integral equation

$$I := \int_0^t U(t-\sigma)\sigma^{-\nu}|u|^{p-1}u(\sigma)d\sigma, \tag{2.1}$$

starting from the Strichartz estimate for the free solution. We use the Hölder inequality and its counterpart for the Besov spaces to estimate the inhomogeneous input term $t^{-\nu}|u|^{p-1}u$. The critical power $p = 1 + 4/(n + 2s)$ was excluded in the preceding works because $t^{-\nu} \notin L^{1/\nu}(0, 1)$. However, if we use the fact $t^{-\nu} \in L^{1/\nu,\infty}$ (Lorentz space), refinements of the inequalities via the real interpolation and well-known fixed point argument in the critical case, we can obtain local wellposedness of (1.5). More specifically, we will use the Lorentz spaces $L^{q,2}$ for t instead of the Lebesgue spaces.

First, we use the real interpolation to refine Hölder's inequality and the Strichartz estimate. We have

$$\|fg\|_{L^{p,q}} \le C\|f\|_{L^{p_0,q_0}}\|g\|_{L^{p_1,q_1}}, \tag{2.2}$$

where $1/p = 1/p_0 + 1/p_1$, $1/q = 1/q_0 + 1/q_1$ and $p, p_0, p_1 < \infty$. Similarly, we have the refinement of Young's inequality (or Hardy–Littlewood–Sobolev)

$$\|f * g\|_{L^{p,q}} \le C\|f\|_{L^{p_0,q_0}} \|g\|_{L^{p_1,q_1}}, \tag{2.3}$$

where $1/p = 1/p_0 + 1/p_1 - 1$, $1/q = 1/q_0 + 1/q_1$, $1 < p, p_0, p_1 < \infty$, and $1 \le q, q_0, q_1$. For convenience, we give the following proof.

Proof. Apply the real interpolation of type (θ, ∞) to Hölder's inequality for f with varying p and p_0. Then we have

$$\|fg\|_{L^{p,\infty}} \le C\|f\|_{L^{p_0,\infty}} \|g\|_{L^{p_1}}. \tag{2.4}$$

Apply the interpolation of type (θ, q_1) to this inequality for g with varying p and p_1. Then we have

$$\|fg\|_{L^{p,q_1}} \le C\|f\|_{L^{p_0,\infty}} \|g\|_{L^{p_1,q_1}}, \tag{2.5}$$

which gives the desired estimate in the case where q_0 or q_1 is infinity. Then, by the complex interpolation for both f and g, we obtain the result for $1 \le q$. The remaining case follows from the power theorem. The proof for the generalized Young is the same. □

If we use (2.3) with $q = q_1 = 2$ and $q_0 = \infty$ instead of the Hardy–Littlewood–Sobolev to prove the Strichartz estimate, we obtain

$$\begin{aligned} \|U(t)\varphi\|_{L^{q,2}(L^r)} &\le C\|\varphi\|_{L^2} \\ \left\| \int_0^t U(t-s) f(s) ds \right\|_{L^{q,2}(L^r)\cap L^\infty(L^2)} &\le C\|f\|_{L^{q',2}(L^{r'})}, \end{aligned} \tag{2.6}$$

where $2 < q < \infty$, $2 \le r \le \infty$, $1/r = 1/2 - 2/(nq)$, and $1/q + 1/q' = 1/r + 1/r' = 1$. Variants of this estimate in the Sobolev or Besov norms for the solutions in H^s are immediate from (2.6) operated by $\mathcal{F}^{-1}\langle x\rangle^s \mathcal{F} = (1-\Delta)^{s/2}$ or the Littlewood–Paley decomposition.

Now it is easy to obtain the following.

THEOREM 2.1. *Let $n \in \mathbb{N}$ and $0 < s < p = 1 + 4/(n+2s) > 1 + 2/n$. Then, the wave operators for NLS are well defined on $\Sigma^s = H^s \cap \mathcal{F}H^s$.*

Proof. It suffices to solve the Cauchy problem (1.5) locally in time, since it is known that the solutions will be automatically global. We can use the fixed point theorem, for example, in the space $S := L^{q,2}(B^s_{r,2})$, where $1/r = 1/2 - 2/(nq)$ and $(p+1)/q + \nu = 1$. By the Sobolev embedding, we have $S \subset X := L^{q,2}(L^\rho)$, where $1/\rho = 1/r - s/n > 0$. By the generalized Hölder inequality (2.2) we have $\|t^{-\nu}|u|^{p-1}u\|_{S'} \le C\|t^{-\nu}\|_{L^{1/\nu,\infty}} \|u\|_X^{p-1} \|u\|_S \le C\|u\|_S^p$, where $S' = L^{q',2}(B^s_{r',2})$ with $1/q' = 1 - 1/q$ and $1/r' = 1 - 1/r$. Then, by the generalized Strichartz estimate (2.6), we obtain $I \in S \cap C(\mathbb{R}; H^s)$. The difference of two solutions can be estimated in the spaces without the derivative. Hence we obtain the unique local solution $u \in C(\mathbb{R}; H^s)$ to the Cauchy problem (1.5) by the standard argument. Next, following [6], we replace $\mathcal{F}^{-1}\langle x\rangle^s \mathcal{F}$ with $U(t)\langle x\rangle^s U(-t)$ in the above argument and use the corresponding Besov-type spaces and commutation properties to deduce that $\langle x\rangle^s U(-t)u(t) \in C(\mathbb{R}; L^2)$ by the same procedure as above. Thus we conclude that $u \in C(\mathbb{R}; \Sigma^s)$. □

REFERENCES

[1] J. E. Barab, *Nonexistence of asymptotically free solutions for a nonlinear Schrödinger equation*, J. Math. Phys., 25 (1984), pp. 3270–3273.

[2] J. Bourgain, *Refinements of Strichartz' inequality and applications to* 2*D-NLS with critical nonlinearity*, Internat. Math. Res. Notices, 5 (1998), pp. 253–283.

[3] J. Bourgain, *Global wellposedness of defocusing critical nonlinear Schrödinger equation in the radial case*, J. Amer. Math. Soc., 12 (1999), pp. 145–171.

[4] T. Cazenave and F. B. Weissler, *Rapidly decaying solutions of the nonlinear Schrödinger equation*, Comm. Math. Phys., 47 (1992), pp. 75–100.

[5] J. Ginibre and T. Ozawa, *Long range scattering for nonlinear Schrödinger and Hartree equations in space dimension* $n \geq 2$, Comm. Math. Phys., 151 (1993), pp. 619–645.

[6] J. Ginibre, T. Ozawa, and G. Velo, *On the existence of the wave operators for a class of nonlinear Schrödinger equations*, Ann. Inst. H. Poincaré Phys. Théor., 60 (1994), pp. 211–239.

[7] J. Ginibre and G. Velo, *On a class of nonliear Schrödinger equations.* II. *Scattering theory, general case*, J. Funct. Anal., 32 (1979), pp. 33-71.

[8] J. Ginibre and G. Velo, *Scattering theory in the energy space for a class of non-linear Schrödinger equations*, J. Math. Pures Appl., 64 (1985), pp. 363–401.

[9] R. T. Glassey, *Asymptotic behavior of solutions to certain nonlinear Schrödinger-Hartree equations*, Comm. Math. Phys., 53 (1977), pp. 9–18.

[10] J. E. Lin and W. A. Strauss, *Decay and scattering of solutions of a nonlinear Schrödinger equation*, J. Funct. Anal., (1978), pp. 245–263.

[11] A. Matsumura, *On the asymptotic behavior of solutions of semi-linear wave equations*, Publ. Res. Inst. Math. Sci., 12 (1976/77), pp. 169–189.

[12] K. Nakanishi, *Energy scattering for nonlinear Klein-Gordon and Schrödinger equations in spatial dimensions* 1 *and* 2, J. Funct. Anal., 169 (1999), pp. 201–225.

[13] T. Ozawa, *Long range scattering for nonlinear Schrödinger equations in one space dimension*, Comm. Math. Phys., 139 (1991), pp. 479–493.

[14] P. Sjölin, *Regularity of solutions to the Schrödinger equations*, Duke Math. J., 55 (1987), pp. 699–715.

[15] W. Strauss, *Nonlinear Wave Equations*, CBMS Reg. Conf. Ser. Math. 73, AMS, Providence, RI, 1989.

[16] Y. Tsutsumi, *Scattering problem for nonlinear Schrödinger equations*, Ann. Inst. H. Poincaré Phys. Théor., 43 (1985), pp. 321–347.

[17] Y. Tsutsumi, *L^2-solutions for nonlinear Schrödinger equations and nonlinear groups*, Funkcial. Ekvac., 30 (1987), pp. 115–125.

[18] Y. Tsutsumi and K. Yajima, *The asymptotic behavior of nonlinear Schrödinger equations*, Bull. Amer. Math. Soc. (N.S.), 11 (1984), pp. 186–188.

SIAM J. MATH. ANAL.
Vol. 32, No. 6, pp. 1272–1310

FAST EVALUATION OF RADIAL BASIS FUNCTIONS: METHODS FOR FOUR-DIMENSIONAL POLYHARMONIC SPLINES*

R. K. BEATSON†, J. B. CHERRIE†, AND D. L. RAGOZIN‡

Abstract. As is now well known for some basic functions ϕ, hierarchical and fast multipole-like methods can greatly reduce the storage and operation counts for fitting and evaluating radial basis functions. In particular, for spline functions of the form

$$s(x) = p(x) + \sum_{k=1}^{N} d_k \phi(|x - x_k|),$$

where p is a low degree polynomial and with certain choices of ϕ, the cost of a single extra evaluation can be reduced from $\mathcal{O}(N)$ to $\mathcal{O}(\log N)$, or even $\mathcal{O}(1)$, operations and the cost of a matrix-vector product (i.e., evaluation at all centers) can be decreased from $\mathcal{O}(N^2)$ to $\mathcal{O}(N \log N)$, or even $\mathcal{O}(N)$, operations.

This paper develops the mathematics required by methods of these types for polyharmonic splines in $\mathbb{R}^4$. That is, for splines s built from a basic function from the list $\phi(r) = r^{-2}$ or $\phi(r) = r^{2n} \ln(r)$, $n = 0, 1, \ldots$. We give appropriate far and near field expansions, together with corresponding error estimates, uniqueness theorems, and translation formulae.

A significant new feature of the current work is the use of arguments based on the action of the group of nonzero quaternions, realized as 2×2 complex matrices

$$\mathbb{H}_0^1 = \left\{ x = \begin{bmatrix} z & w \\ -\overline{w} & \overline{z} \end{bmatrix} : \ |z|^2 + |w|^2 > 0 \right\},$$

acting on $\mathbb{C}^2 = \mathbb{R}^4$. Use of this perspective allows us to give a relatively efficient development of the relevant spherical harmonics and their properties.

Key words. radial basis functions, polyharmonic splines, fast multipole methods

AMS subject classifications. 65D07, 41A15, 41A58, 33C55

PII. S0036141099361767

1. Introduction.

In a large scale comparison of methods for interpolating two-dimensional scattered data Franke [10] identified radial basis functions as one of the most promising methods. These are functions of the form

$$s(\cdot) = p(\cdot) + \sum_{k=1}^{N} d_k \phi(|\cdot - x_k|), \tag{1.1}$$

where p is a low degree polynomial, and the basic function ϕ is usually of noncompact support [17]. Statisticians have also successfully employed radial basis functions fitted by generalized cross validation to smoothing noisy data, e.g., in modeling rainfall distribution across Australia [13]. However, widespread adoption of these techniques has

*Received by the editors September 24, 1999; accepted for publication (in revised form) September 13, 2000; published electronically March 15, 2001.

http://www.siam.org/journals/sima/32-6/36176.html

†Department of Mathematics and Statistics, University of Canterbury, Private Bag 4800, Christchurch, New Zealand (r.beatson@math.canterbury.ac.nz, j.cherrie@math.canterbury.ac.nz). The research of the first and second authors was partially supported by PGSF subcontract DRF601.

‡Department of Mathematics, University of Washington, Box 354350, Seattle, WA 98195 (rag@math.washington.edu). This material is based upon work supported by the National Science Foundation grant DMS-9972004.

been delayed by their apparent extreme computational cost. For example, conventional methods for fitting by interpolation require $\mathcal{O}(N^2)$ storage and $\mathcal{O}(N^3)$ arithmetic operations, where N is the number of data points. This makes computations when N exceeds $10,000$ totally impractical.

These storage and operation counts had led many researchers to incorrectly conclude that the computations are all but impossible. Sibson and Stone [20], talking about problems with $50,000$ to $100,000$ data sites, state: "*We believe such problems will indefinitely remain beyond the scope of thin-plate splines.*" Also Flusser [9], in the context of warping digital images, comments on the computational complexity of evaluation of thin-plate splines, concluding that "*their direct use has extreme computing complexity and is not suitable for practical applications.*"

Recent algorithmic advances involving hierarchical and fast multipole-like methods have invalidated the comments noted above, at least for two- and three-dimensional data; see [3, 4, 7, 8, 11]. These algorithmic developments employ far field and near field expansions to reduce the computational cost of evaluating an N center polyharmonic radial basis function at $m \geq N$ points to $\mathcal{O}(m \log N)$ or even $\mathcal{O}(m)$ operations, at least in $\mathbb{R}^2$ and $\mathbb{R}^3$. Furthermore, interpolatory fitters have been developed which can solve the interpolation problem for such splines in $\mathcal{O}(N)$ storage and $\mathcal{O}(N \log N)$ arithmetic operations. These iterative fitters use the hierarchical evaluators as fast matrix-vector multipliers and can fit interpolatory thin-plate splines to $100,000$ data points in a few minutes on relatively inexpensive workstations.

Polyharmonic splines in $\mathbb{R}^4$ are spline functions of the form (1.1), where the basic function ϕ is a member of the list

$$\phi(r) = \begin{cases} r^{-2}, \\ r^{2n} \ln(r), \quad n = 0, 1, \dots . \end{cases} \tag{1.2}$$

Interpolatory splines of this type minimize suitable energy seminorms, as do their analogues in lower dimensions. For example, the functional

$$I(s) = \int_{\mathbb{R}^4} \sum_{|\alpha|=3} \binom{3}{\alpha} \Big(\big(D^\alpha s\big)(x) \Big)^2 dx$$

is minimized over all suitably smooth functions, satisfying the interpolation conditions

$$s(x_k) = f(x_k), \qquad k = 1, \dots , N, \tag{1.3}$$

if and only if s is the triharmonic spline

$$s(\cdot) = p_2(\cdot) + \sum_{k=1}^{N} d_k |\cdot - x_k|^2 \ln |\cdot - x_k|,$$

where $p_2 \in \pi_2^4$ (the space of quadratics in 4 variables), and the coefficients $\{d_k\}$, and those of the polynomial p_2, are determined by the interpolation conditions (1.3) along with the orthogonality conditions

$$\sum_{k=1}^{N} d_k q(x_k) = 0 \qquad \text{for all } q \in \pi_2^4.$$

Thus polyharmonic splines in $\mathbb{R}^4$ can be expected to be highly successful approximators and interpolators, as experience has shown the polyharmonic splines in lower

dimensions to be. However, meaningful data sets in $\mathbb{R}^4$ can be expected to have many points. Hence, the development of fast evaluation and fitting methods is almost a prerequisite to the use of polyharmonic splines in $\mathbb{R}^4$. Motivated by this we will develop the analytic underpinnings of a fast hierarchical and a fast multipole-like method for polyharmonic splines in $\mathbb{R}^4$. There are many potential applications of these fast methods. One possible application to data mining is estimating the probability of some attribute, such as early death due to heart attack or the filing of a fraudulent tax return, by a regression spline depending on four predictor variables. An application to environmental engineering is modeling the concentration of some chemical, or pollutant, as a function of position and time. Turk and O'Brien [21] suggest using polyharmonic splines in $\mathbb{R}^4$ for shape transformation, or morphing, of implicitly defined three-dimensional surfaces.

The core ideas underlying hierarchical, and fast multipole methods, are beautifully simple. They will be described in the next few paragraphs. First, one needs to accept that only a certain finite precision is necessary. This allows the use of approximations. Second, a suitable far field expansion about 0 must be known for the shifted basic function $\phi(|\cdot - x_<|)$. Here, a far field expansion is a series expansion in which the influence of the source point $x_<$, and the evaluation point x, separates. Furthermore, the series should converge at a geometric rate at all points x with $|x|$ sufficiently large compared to $|x_<|$. Associate with any region T in $\mathbb{R}^d$ the part of the RBF due to sources in T,

$$s_T = \sum_{k:x_k \in T} d_k \phi(|\cdot - x_k|). \tag{1.4}$$

Approximate s_T by a truncated far field expansion r_T, chosen to have appropriate accuracy at all evaluation points x sufficiently far from the center of T. If the number of centers x_k in T is large compared to the number of terms in the far field series, it will be quicker to approximately evaluate $s_T(x)$ by evaluating the series $r_T(x)$ rather than by evaluating $s_T(x)$ directly.

The idea, described above, of using an approximating series when it is faster to do so, lies at the heart of hierarchical methods. Its application is organized by using also a hierarchical subdivision of space. This subdivision determines for a given evaluation point x which parts of s, that is, which s_T's, should be approximated by the corresponding far field series r_T and which parts are to be evaluated directly. In order to be more explicit about the algorithmic organization of the evaluation process, suppose for the moment that space is subdivided into a binary tree of rectangular panels. The root panel will be chosen to contain all the centers, and the children of each parent will be formed by splitting the parent with a hyperplane. Associate with every panel T, the RBF s_T, far field approximation r_T, and a distance from the center of T at which r_T gives a sufficiently accurate approximation to s_T. Then approximate evaluation of $s(x)$ can be performed by a recursive descent of the tree. The actions to be taken when panel T is encountered during this descent are as follows:

- If x is sufficiently far from T, that is, if the distance from x to the center of T is large enough, then the approximation to $s(x)$ is incremented with the approximation $r_T(x)$ to $s_T(x)$.
- Else, if T is childless, the approximation to $s(x)$ is incremented by the directly calculated value of $s_T(x)$.
- Otherwise the process descends to the children of T.

We turn now from algorithmic matters to the analytic underpinnings of a generic

fast multipole method. Results of the following nature are required for the basic function ϕ being used:

- The existence of a rapidly converging far field expansion, centered at 0, for the shifted basic function $\phi(|\cdot - x_<|)$, e.g., Lemmas 4.1 and 4.4.
- Error bounds that determine how many terms are required in each expansion to achieve a specified accuracy, e.g., Theorems 4.2 and 4.10.
- Efficient recurrence relations for computing the coefficients of the expansions, e.g., Lemmas 3.3 and 3.4.
- Uniqueness results that justify indirect translation of expansions, thus allowing the expansions of parent panels to be calculated quickly from those of children, e.g., Lemmas 5.1 and 5.2.
- Formulae for efficiently converting a far field expansion to a rapidly convergent local expansion, e.g., Theorem 6.1.

This paper provides appropriate mathematics for polyharmonic radial basis functions on $\mathbb{R}^4$. That is, for functions of the form (1.1) where ϕ is given by (1.2).

Our discussion above outlines the analytic and algorithmic underpinnings of hierarchical and fast multipole methods. More detailed discussions may be found in the original paper of Greengard and Rokhlin [11], or the introductory short course [3]. Previous papers concerning fast multipole and related methods for fast evaluation of radial basis functions include [4, 5, 6].

A significant technique in our development in this paper is the use of a group action perspective, in particular, of arguments based on the action of the group of nonzero quaternions, realized as 2×2 complex matrices

$$\mathbb{H}_0^1 = \left\{ x = \begin{bmatrix} z & w \\ -\overline{w} & \overline{z} \end{bmatrix} : |z|^2 + |w|^2 > 0 \right\}$$

acting on $\mathbb{C}^2 = \mathbb{R}^4$. We develop almost all the (simple) details needed for these arguments without relying on other presentations of the possibly unfamiliar group representation theory. Use of this perspective allows us to give a relatively efficient development of the relevant spherical harmonics and their properties. See [18, 19] for related analyses of spherical harmonics and their approximation properties. Our work has also been influenced by the elegant and concise treatment due to Epton and Dembart [8] of the analogous expansions for the three-dimensional fast multipole method.

This paper is organized as follows. Section 2 concerns some of the properties of polyharmonic functions on $\mathbb{R}^4$—including realizations of $\mathbb{R}^4$ and representations of $\mathbb{H}_0^1$. It also introduces the inner and outer functions (spherical harmonics) that form the basis of our far field expansions. Section 3 develops a number of properties of these functions that can be applied to far field expansions. These include recurrence formulae, derivative formulae, and symmetries. Section 4 contains the main results on the far field expansions themselves and the associated error bounds. Section 5 develops the uniqueness results that allow the far field expansions to be computed indirectly and economically via the translation theory of section 6. Section 6 also contains the outer-to-inner and inner-to-inner translation formulae needed to approximate far field series by local Taylor series.

2. Polyharmonic functions on $\mathbb{R}^4$. We will represent a nonzero $x \in \mathbb{R}^4$ in three different ways:

$$x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} \quad \text{or} \quad [z, w] \quad \text{or} \quad \begin{bmatrix} z & w \\ -\overline{w} & \overline{z} \end{bmatrix},$$

where $z = x_1 + \mathbf{i}x_2$, $w = x_3 + \mathbf{i}x_4$, and $x_1, \ldots, x_4 \in \mathbb{R}$. The first realization is as an element of $\mathbb{R}^4$, the second is as an element of $\mathbb{C}^2$, and the last as an element of the punctured quaternion (Hamiltonian) space

$$\mathbb{H}_0^1 = \left\{ x = \begin{bmatrix} z & w \\ -\overline{w} & \overline{z} \end{bmatrix} : |z|^2 + |w|^2 > 0 \right\}. \tag{2.1}$$

Note that for elements of $\mathbb{H}_0^1$ the classical adjoint or adjugate and the Hermitian adjoint coincide and

$$x^{-1} = x^* / \det(x).$$

We are primarily interested in $\mathbb{R}^4$ with the usual inner product,

$$\langle x, x_< \rangle = x_1 x_{<1} + x_2 x_{<2} + x_3 x_{<3} + x_4 x_{<4} = |x||x_<| \cos\theta,$$

where $|\cdot|$ is the 2-norm for $\mathbb{R}^4$. In terms of the $\mathbb{C}^2$ realization of $\mathbb{R}^4$ this becomes

$$\langle x, x_< \rangle = \Re(z\overline{z_<} + w\overline{w_<}) = \tfrac{1}{2}(z\overline{z_<} + w\overline{w_<} + z_<\overline{z} + w_<\overline{w})$$

and, in terms of the matrix realization $\mathbb{H}_0^1$,

$$\langle x, x_< \rangle = \tfrac{1}{2}\operatorname{Tr}(x^* x_<) = \tfrac{1}{2}|x|^2 \operatorname{Tr}(x^{-1} x_<) = \tfrac{1}{2}\operatorname{Tr}({x_<}^* x),$$

where Tr is the trace. Note that this inner product gives the norm

$$|x|^2 = x_1^2 + x_2^2 + x_3^2 + x_4^2 = z\overline{z} + w\overline{w} = \det(x).$$

We also require an inner product for functions on the unit ball $B = \{x \in \mathbb{R}^4 : |x| \le 1\}$. For $f, g \in L^2(B)$ we define their inner product by

$$(f, g) = \int_{|\xi| \le 1} f(\xi)\overline{g(\xi)} d\xi. \tag{2.2}$$

We will also use this pairing for other functions f and g on B with $f\overline{g} \in L^1(B)$. Furthermore, we will also require the subspaces of $C(\mathbb{R}^4)$ defined by

$$\mathcal{H}_m = \{p([z_1, z_2]) : p \text{ is a homogeneous polynomial of degree } m \text{ in } z_1, z_2\},$$

where $m \in \mathbb{N}_0$, the set of nonnegative integers.

2.1. Irreducible representations of $\mathbb{H}_0^1$-spherical harmonics. In this section we will develop irreducible representations of $\mathbb{H}_0^1$ in $\mathcal{H}_m$. Our purpose for doing so is that the coefficient functions of these representations will eventually be seen to form a very computationally convenient basis for the harmonic polynomials in $\mathbb{R}^4$. In fact, when these coefficient functions are multiplied by $|x|^{2\ell}, \ell = 0, \ldots, k$, they yield

a basis for all $(k+1)$-harmonic polynomials in $\mathbb{R}^4$. We note in particular the simple form of the addition formulae (Lemma 2.9), the recurrence relation (Lemma 3.3), and the dual basis (Lemma 3.10), to come.

Most of the relevant representation theory and mathematical physics literature is focused on rotations of $\mathcal{S}^2$ or $\mathcal{S}^3$ and therefore considers representations of

$$SU(2) = \left\{ x \in \mathbb{H}_0^1 \,:\, \det(x) = |x|^2 = 1 \right\} = \mathcal{S}^3.$$

However, in the context of far field expansions it is convenient to work instead with all of $\mathbb{H}_0^1$ to take into account both the scaling by $|x|$ and rotation by elements of $SU(2)$. This leads to some differences, most importantly the character functions now depend on the norm of x as well as the angle θ between x and the north pole $[1,0]$. Other differences include the formula for the product of two character functions and the addition formulae.

DEFINITION 2.1. *Given a group G, a* representation $T : G \to \mathrm{GL}(V)$ *of G on V is an operator valued map that satisfies $T(g \cdot h) = T(g)T(h)$. A representation T of G on V is* irreducible *if the only subspaces of V that are invariant under $T(g)$ for all $g \in G$ are $\{0\}$ and V.*

We define representations $T_m(x)$ of $\mathbb{H}_0^1$ in the spaces $\mathcal{H}_m$ given by

$$\begin{aligned} T_m(x)\; p([z_1, z_2]) &= p([z_1, z_2]x) \\ &= p([z_1 z - z_2 \overline{w},\; z_1 w + z_2 \overline{z}]). \end{aligned} \tag{2.3}$$

Note that since $\mathcal{H}_m$ is embedded in the space of functions on $\mathbb{H}^1$ (or even on $\mathbb{H}_0^1$), this representation is just the restriction of the right action defined by

$$\big(x \cdot f\big)([z_1, z_2]) = f([z_1, z_2]x), \quad x \in \mathbb{H}_0^1.$$

If we put a (Hilbert space) norm on these functions via (2.2), i.e.,

$$\|f\|^2 = (f, f) = \int_{|\xi| \le 1} |f(\xi)|^2 d\xi,$$

then the rotation invariance of Lebesgue measure implies that $\|x \cdot f\| = \|f\|$, whenever $|x| = 1$, i.e., whenever $x \in \mathcal{S}^3$. Thus

$$\big(T_m(x)f, T_m(x)f\big) = (f, f)$$

for all functions $f \in L^2(B)$, and $T_m(x)$, as an operator, is unitary if $|x| = 1$. The reader is cautioned that the matrix realization of $T_m(x)$ to come is not unitary, but can be scaled to be so (see Lemma 2.5).

LEMMA 2.2. *The representations* (2.3) *are irreducible.*

Proof. Assume there is a subspace $V \subset \mathcal{H}_m$ that is invariant under $T_m(x)$ for all $x \in \mathbb{H}_0^1$. Then V is invariant under $T_m(x)$ if we restrict attention to $x \in SU(2) \subset \mathbb{H}_0^1$. Since the representations T_m of $SU(2)$ are irreducible [16, pp. 208–211], it follows that either $V = \{0\}$ or $V = \mathcal{H}_m$ and hence that the representations T_m are irreducible. ∎

The monomials

$$e_k^m([z_1, z_2]) = z_1^{m-k} z_2^k, \qquad 0 \le k \le m,$$

form a basis for $\mathcal{H}_m$. The operators $T_m(x)$ have a matrix realization once this basis for $\mathcal{H}_m$ is chosen. The elements of these matrices will be denoted $t^m_{i,j}$ and are given by

$$\begin{aligned} t^m_{i,j}(x) &= t^m_{i,j}(z,w) \\ &= \text{coefficient of } e^m_i \text{ in } T_m(x)e^m_j \\ &= \text{coefficient of } z_1^{m-i}z_2^i \text{ in } (z_1z - z_2\overline{w})^{m-j}(z_1w + z_2\overline{z})^j. \end{aligned}$$

Equivalently,

$$\sum_{i=0}^{m} t^m_{i,j}(z,w)z_1^{m-i}z_2^i = (z_1z - z_2\overline{w})^{m-j}(z_1w + z_2\overline{z})^j. \tag{2.4}$$

For $m = 0$, $t^0_{0,0}(x) = 1$, while for $m = 1$, from (2.4)

$$\begin{aligned} t^1_{0,0}(z,w) &= z, & t^1_{0,1}(z,w) &= w, \\ t^1_{1,0}(z,w) &= -\overline{w}, & t^1_{1,1}(z,w) &= \overline{z}, \end{aligned}$$

or in matrix terms

$$[T_1(x)] = x. \tag{2.5}$$

An immediate consequence of this choice of basis is the following lemma.

LEMMA 2.3. *Treated as matrices, $T_m(z,0)$ is a diagonal matrix and $T_m(0,w)$ is an antidiagonal matrix. Specifically these matrices have entries*

$$t^m_{i,j}(z,0) = \begin{cases} 0, & i \neq j, \\ z^{m-i}\overline{z}^i, & i = j, \end{cases}$$

and

$$t^m_{i,j}(0,w) = \begin{cases} 0, & i \neq m-j, \\ w^{m-i}(-\overline{w})^i, & i = m-j. \end{cases}$$

The basis elements e^m_k for $\mathcal{H}_m$ are orthogonal with respect to the inner product (2.2). In fact the exact norms for the basis elements e^m_k and thus the row and column scalings to get unitary matrices are easily computed.

LEMMA 2.4. *The basis functions e^m_k are orthogonal with inner products given by*

$$\left(e^m_k, e^{m'}_{k'}\right) = \int_{|\xi|\leq 1} e^m_k(\xi)\overline{e^{m'}_{k'}(\xi)}d\xi = \delta_{m,m'}\delta_{k,k'}\pi^2\frac{k!(m-k)!}{(m+2)!}.$$

Proof. Introduce polar coordinates (r_1,θ_1) and (r_2,θ_2) in the z_1 and z_2 planes (where $\xi = [z_1, z_2]$).

$$\begin{aligned} \left(e^m_k, e^{m'}_{k'}\right) &= \int_{|\xi|\leq 1} e^m_k(\xi)\overline{e^{m'}_{k'}(\xi)}d\xi \\ &= \int_0^1 \int_0^{\sqrt{1-r_1^2}} \int_0^{2\pi} \int_0^{2\pi} r_1^{m-k}e^{\mathbf{i}(m-k)\theta_1}r_2^k e^{\mathbf{i}k\theta_2} \end{aligned}$$

$$\begin{aligned}
&\times r_1^{m'-k'} e^{-\mathbf{i}(m'-k')\theta_1} r_2^{k'} e^{-\mathbf{i}k'\theta_2}\, d\theta_2 d\theta_1 r_2 dr_2 r_1 dr_1 \\
&= (2\pi)^2 \delta_{k,k'}\delta_{m,m'} \int_0^1 \int_0^{\sqrt{1-r_1^2}} r_1^{2(m-k)+1} r_2^{2k+1} dr_2 dr_1 \\
&= (2\pi)^2 \delta_{k,k'}\delta_{m,m'} \int_0^1 r_1^{2(m-k)+1} \frac{(1-r_1^2)^{k+1}}{2k+2} dr_1 \\
&= \pi^2 \delta_{k,k'}\delta_{m,m'} \mathrm{B}(m-k+1,k+2)/(k+1) \\
&= \pi^2 \delta_{k,k'}\delta_{m,m'} \frac{(m-k)!(k+1)!}{(m+2)!(k+1)},
\end{aligned}$$

where B is the Beta function $\mathrm{B}(n,m) = \Gamma(n)\Gamma(m)/\Gamma(n+m)$. ∎

Since T_m restricted to $\mathcal{S}^3$ acts in a norm preserving way on $\mathcal{H}_m$, we easily obtain the following matrix representation for $T_m(x^{-1})$.

LEMMA 2.5. *There exist row and column scalings that make the matrices $T_m(x)$ unitary for $|x| = 1$. Specifically*

(i) *The inverse of $T_m(x)$ is given by*

$$T_m(x^{-1}) = \left[t_{i,j}^m(x^{-1})\right] = \left[|x|^{-2m}\overline{t_{j,i}^m(x)}\binom{m}{i}\binom{m}{j}^{-1}\right],$$

or equivalently via

$$t_{i,j}^m(x^*) = t_{i,j}^m(\overline{z},-w) = \overline{t_{j,i}^m(z,w)}\binom{m}{i}\binom{m}{j}^{-1}.$$

(ii) *For all $x \neq 0$, the inverses of the matrices*

$$U_m(x) = \left[t_{i,j}^m(x)\sqrt{\binom{m}{j}\binom{m}{i}^{-1}}\right]$$

are given by $U_m(x)^{-1} = |x|^{-2m}U_m(x)^*$ and thus these matrices are unitary when $|x| = 1$.

Proof. The definition of $t_{j,i}^m$ and the orthogonality of $\{e_k^m : k = 0,\dots,m\}$ implies

$$\left(e_j^m, T_m(x)e_i^m\right) = \left(e_j^m, t_{j,i}^m(x)e_j^m\right) = \overline{t_{j,i}^m(x)}(e_j^m, e_j^m). \tag{2.6}$$

Since $T_m\big((x/|x|)^{-1}\big)$ preserves the inner product (2.2) and is homogeneous of degree m,

$$\begin{aligned}
\left(e_j^m, T_m(x)e_i^m\right) &= \Big(T_m\big((x/|x|)^{-1}\big)e_j^m, T_m\big((x/|x|)^{-1}\big)T_m(x)e_i^m\Big) \\
&= \left(|x|^m T_m(x^{-1})e_j^m, |x|^m e_i^m\right) = |x|^{2m} t_{i,j}^m(x^{-1})(e_i^m, e_i^m).
\end{aligned} \tag{2.7}$$

Equating (2.6) to (2.7) and solving, we obtain

$$t_{i,j}^m(x^{-1}) = |x|^{-2m}\overline{t_{j,i}^m(x)}\frac{(e_j^m, e_j^m)}{(e_i^m, e_i^m)}.$$

Taking into account the previous lemma and the fact that $x^{-1} = [\overline{z},-w]/(z\overline{z}+w\overline{w})$, this gives the desired results. ∎

Since it is easy to use the definitions to show the first row of each $T_m(x)$ is given by

$$t^m_{0,j}(x) = t^m_{0,j}(z,w) = z^{m-j}w^j = e^m_j([z,w]),$$

part of Lemma 2.4 shows that $\{t^m_{0,j}\}$ is orthogonal with respect to the inner product (2.2). In fact much more general (bi-) orthogonality facts are true for $t^m_{i,j}(x)$ and $\overline{t^m_{j,i}(x^{-1})}$. These are related to the orthogonality properties of the irreducible unitary matrix representations of any compact group, such as $\mathcal{S}^3$, as in [12, (27.19)]. But we prefer to present them in a slightly more general form which is closely related to the coordinate-free proofs in Chapter 3 of [1], particularly Proposition 3.15, Schur's Lemma 3.22, and its corollary, 3.23.

LEMMA 2.6.

(i) *(Schur's lemma) For any $(m+1)\times(m+1)$ matrix A*

$$\widetilde{A} = \int_{0<|x|\le 1} T_m(x^{-1})AT_m(x)dx = cI,$$

where

$$c = \frac{\mathrm{vol}\{|x|\le 1\}}{m+1}\,\mathrm{Tr}(A) = \frac{\pi^2/2}{m+1}\,\mathrm{Tr}(A).$$

(ii) *The set*

$$\left\{ \frac{(m+1)}{(\pi^2/2)}\binom{m}{j}\binom{m}{i}^{-1} |\cdot|^{-2m} t^m_{i,j}(\cdot) = \frac{m+1}{\pi^2/2}\overline{t^m_{j,i}(\cdot^{-1})}, i,j = 0,\dots,m \right\}$$

is biorthogonally dual to $\{t^m_{i,j}(\cdot), i,j = 0,\dots,m\}$ with respect to the pairing (2.2). That is,

$$\int_{0<|x|\le 1} \frac{m+1}{\pi^2/2} t^m_{i',j'}(x)t^m_{j,i}(x^{-1})dx = \delta_{i,i'}\delta_{j,j'}.$$

(iii) *The first two parts are also true when $\{0<|x|\le 1\}$ is replaced by $\mathcal{S}^3$ and "vol" is replaced by "surface area" (so $\pi^2/2$ is replaced by $2\pi^2$).*

Proof. For (i), let $y\in\mathcal{S}^3$ be arbitrary. Then

(2.8)
$$\widetilde{A}T_m(y) = \int_{0<|x|\le 1} T_m(x^{-1})AT_m(xy)dx = \int_{0<|x|\le 1} T_m(yx^{-1})AT_m(x)dx = T_m(y)\widetilde{A},$$

since $x\mapsto xy^{-1}$ leaves Lebesgue measure invariant. Let c be any eigenvalue for $\widetilde{A}$ with v an associated eigenvector. From (2.8) $T_m(y)v$ is also an eigenvector for the same eigenvalue c. By the irreducibility of T_m, $\mathrm{span}\{T_m(y)v : y\in\mathcal{S}^3\} = \mathcal{H}_m$. Thus $\widetilde{A}v = cv$ for all vectors v and it follows that $\widetilde{A} = cI$.

To get the formula for c, take the trace of all terms. Then move the linear functional Tr inside the integral and use $\mathrm{Tr}(T_m(x^{-1})AT_m(x)) = \mathrm{Tr}(A)$ to obtain

$$\mathrm{Tr}(cI) = (m+1)c = \int_{0<|x|\le 1} \mathrm{Tr}(A)dx = \mathrm{vol}\{0<|x|\le 1\}\,\mathrm{Tr}(A).$$

For (ii) substitute $A = E_{i,i'} = [\delta_{i,j}\delta_{j',i'}]$ into (i), i.e., use the matrix A with 1 at row i, column i' and zero elsewhere and then use j, j' to index the matrix. Then

$$T_m(x^{-1})E_{i,i'}T_m(x) = [t^m_{j,i}(x^{-1})t^m_{i',j'}(x)].$$

Since $\mathrm{Tr}(E_{i,i'}) = \delta_{i,i'}$, (i) yields

$$\left[\int_{0<|x|\le 1} t^m_{j,i}(x^{-1})t^m_{i',j'}(x)dx\right] = \frac{\pi^2/2}{m+1}\delta_{i,i'}[\delta_{j,j'}].$$

For (iii) repeat the proofs with the ball replaced by the sphere. Or, more simply, just note that if $d\Omega$ denotes the standard "surface" measure on S^3, then integration in spherical coordinates is with respect to $|x|^3 d\Omega(x/|x|)d|x|$. Then, due to the homogeneity of T_m, (i) becomes

$$\int_{\mathcal{S}^3}\int_{|x|=0}^{|x|=1} T_m\big((x/|x|)^{-1}\big)AT_m(x/|x|)|x|^3 d|x|d\Omega(x/|x|)$$
$$= \frac{1}{4}\int_{\mathcal{S}^3} T_m\big((x/|x|)^{-1}\big)AT_m(x/|x|)d\Omega(x/|x|) = \frac{\pi^2/2}{m+1}\,\mathrm{Tr}(A).$$

Clearing the denominator of 4 leads to the desired formula for (i) on $\mathcal{S}^3$. Now (ii) on $\mathcal{S}^3$ follows by exactly the same reasoning. Since $|x|^{-2m} = 1$ on $\mathcal{S}^3$, the orthogonality of $\{t^m_{i,j}\}$ on $\mathcal{S}^3$ follows, as does their independence since none of these functions are zero (or have norm zero) on $\mathcal{S}^3$. □

Lemma 2.6(iii) implies that the coefficient functions are linearly independent in both $L^2(\mathcal{S}^3)$ and $C(\mathbb{R}^4)$. Indeed, they form a basis for homogeneous harmonic polynomials on $\mathbb{R}^4$ and for the spherical harmonics of degree m on $\mathcal{S}^3$ (see (3.14)). Hence, for any $p \in \mathbb{N}$, $\{t^m_{i,j} : 0 \le i, j \le m, 0 \le m \le p\}$ is linearly independent both on $\mathcal{S}^3$ and on $\mathbb{R}^4$.

Given the north pole $[1, 0]$ and some general vector $x = [z, w]$, we can find a rotation that leaves the north pole fixed and rotates x to a vector in the direction $[e^{\mathbf{i}\theta}, 0]$, where $\cos\theta = \Re(z)/|x|$. Note that θ is just the angle between x and the north pole. We could equivalently rotate to a vector in the direction $[e^{-\mathbf{i}\theta}, 0]$. Hence any function independent of rotation about the north pole must be a function of $|x|$ and θ and furthermore must be even in θ, i.e., is a function of $\cos\theta$. It is known that all rotations leaving the north pole fixed can be achieved by conjugation, $x \mapsto vxv^{-1}$, with elements of $SU(2)$. See [14, pp. 277–279] or [2, pp. 214–217] for a geometric proof.

The same result may be obtained algebraically by considering the diagonalizability of x (see, e.g., [16, pp. 209–210]). Therefore there is a $v \in SU(2)$ such that

$$x = v\gamma v^{-1},$$

where

$$\gamma = |x|\,\mathrm{diag}(e^{\mathbf{i}\theta}, e^{-\mathbf{i}\theta}).$$

Note that conjugation with $-v$ achieves the same rotation as conjugation with v, but this is the only nonuniqueness in identifying conjugations with rotations of the equatorial $\mathcal{S}^2$.

These conjugation facts lead to explicit formulae for the traces of these representations. Specifically,

$$\begin{aligned}\operatorname{Tr}\big(T_m(x)\big) = \operatorname{Tr}\big(T_m(v\gamma v^{-1})\big) &= \operatorname{Tr}\big(T_m(v)T_m(\gamma)T_m(v^{-1})\big)\\ &= \operatorname{Tr}\big(T_m(v)T_m(\gamma)T_m(v)^{-1}\big) = \operatorname{Tr}\big(T_m(\gamma)\big).\end{aligned}$$

Using Lemma 2.3,

$$t^m_{i,j}(\gamma) = t^m_{i,j}(|x|e^{\mathbf{i}\theta}) = \begin{cases} |x|^m e^{\mathbf{i}(m-2j)\theta}, & i = j,\\ 0, & i \neq j.\end{cases}$$

Hence

$$\begin{aligned}\operatorname{Tr}\big(T_m(x)\big) = \operatorname{Tr}\big(T_m(\gamma)\big) &= \sum_{j=0}^{m} t^m_{j,j}(\gamma)\\ &= \sum_{j=0}^{m} |\gamma|^m e^{\mathbf{i}(m-2j)\theta}\\ &= |\gamma|^m \frac{e^{\mathbf{i}(m+1)\theta} - e^{-\mathbf{i}(m+1)\theta}}{e^{\mathbf{i}\theta} - e^{-\mathbf{i}\theta}} \qquad \text{if } \theta \neq 0,\\ &= |x|^m \frac{\sin(m+1)\theta}{\sin\theta},\end{aligned}$$

and interpreting $\sin(m+1)\theta/\sin\theta$ in the conventional fashion (as $m+1$) at $\theta = 0$, the expression is also valid there. As usual the character of the representation is defined to be the function $\chi_m : \mathbb{H}^1_0 \to \mathbb{R}$ given by the trace

$$\chi_m(x) := \operatorname{Tr}\big(T_m(x)\big) = |x|^m \frac{\sin(m+1)\theta}{\sin\theta}. \tag{2.9}$$

In particular $\chi_0(x) = 1$ and $\chi_1(x) = \operatorname{Tr}(x) = 2|x|\cos\theta$. Since the entries in $T_m(x)$ are homogeneous polynomials of degree m in z, w, $\overline{w}$, and $\overline{z}$, the χ_m are also. We extend the definition of χ_m to $x = 0$ by continuity and define $\chi_{-1} = 0$. Note that these χ_m are multiples of the Chebyshev polynomials of the second kind as functions of $t = \cos\theta$.

LEMMA 2.7. *For $x \in \mathbb{H}^1$ and $m \in \mathbb{N}_0$,*

$$\chi_1(x)\chi_m(x) = \chi_{m+1}(x) + |x|^2\chi_{m-1}(x). \tag{2.10}$$

Proof. The result is trivially true when $m = 0$ or $x = 0$. For $m > 0$ and $x \neq 0$

$$\begin{aligned}\chi_1(x)\chi_m(x) &= \frac{|x|^{m+1}}{\sin^2\theta}\Big\{\sin(2\theta)\sin\big((m+1)\theta\big)\Big\}\\ &= \frac{|x|^{m+1}}{\sin^2\theta}\Big\{2\sin(\theta)\cos(\theta)\sin\big((m+1)\theta\big)\Big\}\\ &= \frac{|x|^{m+1}}{\sin\theta}\Big\{\sin\big((m+2)\theta\big) + \sin(m\theta)\Big\}\\ &= \chi_{m+1}(x) + |x|^2\chi_{m-1}(x). \qquad \square\end{aligned}$$

2.2. Inner and outer functions. We will refer to the functions $t^m_{i,j}$ of the previous subsection as the *inner* functions as they will be shown to be homogeneous of nonnegative degree and harmonic in $\mathbb{R}^4$. The purpose of this subsection is to introduce the *outer* functions $o^m_{i,j}$ which will be shown to be homogeneous of negative degree and harmonic in $\mathbb{R}^4\backslash\{0\}$. The subsection also contains the addition formula connecting the inner and outer functions with the character functions.

The representations T_m may be used to construct antirepresentations, O_m of $\mathbb{H}^1_0$, defined by

$$O_m(x) := |x|^{-2} T_m(x^{-1}) \tag{2.11}$$

or, in terms of the coefficient functions,

$$\begin{aligned} o^m_{i,j}(z,w) = |x|^{-2} t^m_{i,j}(x^{-1}) &= (z\overline{z} + w\overline{w})^{-1} t^m_{i,j}(\overline{z}/|x|^2, -w/|x|^2) \\ &= |x|^{-(2m+2)} t^m_{i,j}(\overline{z}, -w) \\ &= |x|^{-2(m+1)} \overline{t^m_{j,i}(z,w)} \binom{m}{i} \binom{m}{j}^{-1}. \end{aligned} \tag{2.12}$$

Thus the coefficient functions for O_m are homogeneous of degree $-(m+2)$. Together, (2.4) and (2.12) give the equivalent definition of the outer functions

$$|x|^{2(m+1)} \sum_{i=0}^{m} o^m_{i,j}(z,w) z_1^{m-i} z_2^i = (z_1\overline{z} + z_2\overline{w})^{m-j} \big(z_1(-w) + z_2 z\big)^j. \tag{2.13}$$

See the appendix for tables of low degree inner and outer functions. These O_m are antirepresentations as

$$O_m(x \cdot y) = O_m(y) \cdot O_m(x).$$

Remark 2.8. When we prove harmonicity of the inner functions, or of the outer functions, the harmonicity of the other set will follow. Indeed, definition (2.12) corresponds to an inversion of the functions $t^m_{i,j}$ in the unit sphere followed by scaling by $|x|^{-2}$, along with reflection in the $\Re(z)$ axis. This reflection $(z,w) \to (\overline{z}, -w)$ corresponds to quaternionic conjugation. Both this scaled inversion, sometimes called the Kelvin transformation, and the reflection preserve harmonicity.

Associated with the spherical harmonics in $\mathcal{S}^{d-1}$ for any integer $d \geq 2$ is an addition formula [15, pp. 3–10]. These addition formulae express the character function (sometimes called the Gengenbauer polynomial or the Legendre function) at the inner product of two points u and v on $\mathcal{S}^{d-1}$ as a sum of products, in each of which the influence of u and v is separated. Perhaps the best known example of this phenomena is the addition formula for the ordinary Legendre polynomial, $P_n(\cos\gamma)$, which is exploited in the multipole expansion of the three-dimensional potential (see, e.g., [3, 7, 8]). With our definition of the inner and outer functions, the addition formula for $\mathbb{R}^4$ takes the following forms.

LEMMA 2.9 (Addition formulae for χ_m).

(i) *If* $x, x_< \in \mathbb{H}^1$, $x \neq 0$, *then*

$$\chi_m(x^{-1}x_<) = |x|^2 \operatorname{Tr}\big(O_m(x) T_m(x_<)\big) = |x|^2 \sum_{i,j=0}^{m} t^m_{j,i}(x_<)\, o^m_{i,j}(x). \tag{2.14}$$

(ii) *If* $x, x_< \in \mathbb{H}^1$, *then*

$$\chi_m(x^* x_<) = \mathrm{Tr}\left(T_m(x_<{}^*) T_m(x)\right) = \sum_{i,j=0}^{m} t_{j,i}^m(x_<{}^*)\, t_{i,j}^m(x). \tag{2.15}$$

Proof. From the definition of the character functions χ_m,

$$\begin{aligned}
\chi_m(x^{-1}x_<) = \mathrm{Tr}\left(T_m(x^{-1}x_<)\right) &= \mathrm{Tr}\left(T_m(x^{-1})T_m(x_<)\right) \\
&= \sum_{i,j=0}^{m} t_{i,j}^m(x^{-1}) t_{j,i}^m(x_<) \\
&= |x|^2 \sum_{i,j=0}^{m} |x|^{-2} t_{i,j}^m(x^{-1}) t_{j,i}^m(x_<) \\
&= |x|^2 \sum_{i,j=0}^{m} o_{i,j}^m(x) t_{j,i}^m(x_<),
\end{aligned}$$

which proves part (i). For part (ii), notice that by (2.9),

$$\chi_m(x^*) = \chi_m(x).$$

Therefore

$$\chi_m(x^* x_<) = \chi_m(x_<{}^* x) = \mathrm{Tr}\left(T_m(x_<{}^*)T_m(x)\right) = \sum_{i,j=0}^{m} t_{j,i}^m(x_<{}^*)\, t_{i,j}^m(x). \qquad \square$$

The addition formula (2.14) essentially displays the fact that

$$\frac{m+1}{\pi^2/2}\chi_m(x^{-1}x_<)$$

is a reproducing kernel for $\mathrm{span}\{t_{i,j}^m, i,j = 0, \ldots, m\}$, the space of homogeneous harmonic polynomials of degree m. In fact the biorthogonality in Lemma 2.6(ii) immediately shows that for any

$$f_m = \sum_{i,j=0}^{m} a_{i,j} t_{i,j}^m$$

in this span,

$$\int_{0<|x|\le 1} f_m(x) \frac{m+1}{\pi^2/2} \chi_m(x^{-1}x_<) dx = f_m(x_<).$$

3. Properties of the inner and outer functions. Some fundamental properties of the inner and outer functions will be developed in this section. These properties are needed for the development of a fast multipole-like method, but are also of interest in their own right. Properties which are considered below include symmetries, recurrence relations, derivative formulae, and harmonicity.

3.1. Symmetries. In this subsection we will develop some symmetry properties of the inner and outer functions. We have already seen an example of a symmetry relation in Lemma 2.5. One application of these symmetry properties is an approximate halving of the costs of forming and evaluating the truncated far field expansions to be developed in section 4.

We will use the symbols i, j, and k for the fundamental quaternionic units and $\mathbf{i}$ for the imaginary number $\sqrt{-1}$. It will be convenient to use the 2×2 matrix realization of the quaternions. In this realization the quaternion 1 is the 2×2 identity matrix and

$$\mathsf{i} = \begin{bmatrix} \mathbf{i} & 0 \\ 0 & -\mathbf{i} \end{bmatrix}, \qquad \mathsf{j} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \qquad \mathsf{k} = \begin{bmatrix} 0 & \mathbf{i} \\ -\mathbf{i} & 0 \end{bmatrix}.$$

Now consider conjugation of an element in $\mathbb{H}^1$ by i.

$$\begin{bmatrix} \mathbf{i} & 0 \\ 0 & -\mathbf{i} \end{bmatrix} \begin{bmatrix} z & w \\ -\overline{w} & \overline{z} \end{bmatrix} \begin{bmatrix} -\mathbf{i} & 0 \\ 0 & \mathbf{i} \end{bmatrix} = \begin{bmatrix} z & -w \\ \overline{w} & \overline{z} \end{bmatrix}.$$

Applying T_m

$$T_m(\mathbf{i}, 0) T_m(z, w) T_m(-\mathbf{i}, 0) = T_m(z, -w).$$

By Lemma 2.3 $T_m(\mathbf{i}, 0)$ and $T_m(-\mathbf{i}, 0)$ are diagonal with

$$t^m_{i,i}(\mathbf{i}, 0) = (-1)^i \mathbf{i}^m, \qquad t^m_{j,j}(-\mathbf{i}, 0) = (-1)^{m-j} \mathbf{i}^m.$$

We conclude that

$$(-1)^i \mathbf{i}^m t^m_{i,j}(z, w) (-1)^{m-j} \mathbf{i}^m = t^m_{i,j}(z, -w)$$

for $0 \le i, j \le m$. That is,

$$t^m_{i,j}(z, w) = (-1)^{i-j} t^m_{i,j}(z, -w) \tag{3.1}$$

for $0 \le i, j \le m$.

Similar results can be obtained in the same way by conjugation with j and k. These results are summarized in the following lemma.

LEMMA 3.1. *The inner functions $t^m_{i,j}$ defined by* (2.4) *satisfy*

$$t^m_{i,j}(z, w) = (-1)^{i-j} t^m_{i,j}(z, -w), \tag{3.2a}$$

$$t^m_{i,j}(z, w) = (-1)^{i+j} t^m_{m-i,m-j}(\overline{z}, \overline{w}), \tag{3.2b}$$

$$t^m_{i,j}(z, w) = t^m_{m-i,m-j}(\overline{z}, -\overline{w}) \tag{3.2c}$$

for all $x = [z, w] \in \mathbb{C}^2$.

Because each $t^m_{i,j}(z, w)$ is a polynomial with real coefficients in z, w, $\overline{z}$, and $\overline{w}$,

$$\overline{T_m(z, w)} = T_m(\overline{z}, \overline{w}).$$

Hence (3.2b) may be written

$$t^m_{i,j}(z, w) = (-1)^{i+j} \overline{t^m_{m-i,m-j}(z, w)}. \tag{3.3}$$

Using the expression (2.12) for the outer functions in terms of the inner functions, each symmetry of T_m implies a symmetry of O_m. In particular, symmetry (3.3) implies

$$\begin{aligned} o^m_{i,j}(z,w) &= |x|^{-(2m+2)} t^m_{i,j}(\overline{z},-w) \\ &= |x|^{-(2m+2)}(-1)^{i+j} t^m_{m-i,m-j}(z,-\overline{w}) \\ &= (-1)^{i+j}\overline{o^m_{m-i,m-j}(z,w)}. \end{aligned} \tag{3.4}$$

From symmetries (3.3) and (3.4)

$$o^m_{i,j}(x)t^m_{j,i}(x_<) = \overline{o^m_{m-i,m-j}(x)t^m_{m-j,m-i}(x_<)}$$

for $0 \le i,j \le m$. Hence the terms in the addition formula expression (2.14) for $\chi_m(x^{-1}x_<)$ are conjugate symmetric with respect to reflection in the middle index $(m/2, m/2)$. Thus the addition formula can be rewritten to involve a real part and approximately half the number of terms. For example, one such recasting is

$$\begin{aligned} \chi_m(x^{-1}x_<) &= |x|^2 \sum_{i,j=0}^{m} o^m_{i,j}(x)t^m_{j,i}(x_<) \\ &= |x|^2 \Re\left(\sum_{k=0}^{\left\lfloor \frac{1}{2}\left((m+1)^2-1\right)\right\rfloor} (2-\delta_{m/2,i(k)}\,\delta_{m/2,j(k)}) o^m_{i(k),j(k)}(x)t^m_{j(k),i(k)}(x_<) \right), \end{aligned} \tag{3.5}$$

where $i(k) = k \mod (m+1)$ and $j(k) = \lfloor k/(m+1)\rfloor$. This observation can and should be used to halve the storage requirements and flop counts of fast evaluators built upon the analysis of this paper. For example, the recast expression (3.5) can be substituted almost directly into the truncated far field expansion, g_p, for a sum of shifts of the potential of Theorem 4.2. This will then approximately halve the marginal operation count for evaluating $g_p(x)$ at a single extra x.

A further symmetry will be useful to recast products of powers of $|x|$ with outer functions as inner functions.

LEMMA 3.2. *For $m \in \mathbb{N}$ and $0 \le i,j \le m$*

$$\binom{m}{j}|x|^{2m+2}o^m_{i,j}(x) = (-1)^{i-j}\binom{m}{i}t^m_{m-j,m-i}(x).$$

Proof. Substitution of (3.3) into (2.12) gives the result. ∎

3.2. Recurrence relationships. In this subsection we develop recurrence relations which provide efficient methods for calculating the inner and outer functions. For a given x they allow calculation of $\{t^m_{i,j},\ 0 \le i,j \le m,\ 0 \le m \le p\}$ or $\{o^m_{i,j},\ 0 \le i,j \le m,\ 0 \le m \le p\}$ in $\mathcal{O}(p^3)$ operations, which is the same magnitude as the number of terms to be calculated.

It is convenient to extend the definitions of $t^m_{i,j}$ and $o^m_{i,j}$ by defining $t^m_{i,j} = 0 = o^m_{i,j}$ for i or $j \notin [0,m]$.

LEMMA 3.3. *The inner functions $t^m_{i,j}$ defined by* (2.4) *satisfy the overlapping recurrence relations*

$$t^{m+1}_{i,j}(z,w) = \begin{cases} z\, t^m_{i,j}(z,w) - \overline{w}\, t^m_{i-1,j}(z,w), & 0 \le j \le m, \\ w\, t^m_{i,j-1}(z,w) + \overline{z}\, t^m_{i-1,j-1}(z,w), & 1 \le j \le m+1 \end{cases} \tag{3.6}$$

for $0 \le i \le m+1$ *and also satisfy the backward recurrence relations*

$$
\begin{aligned}
|x|^2 t^m_{i,j}(z,w) &= \overline{z}\; t^{m+1}_{i,j}(z,w) + \overline{w}\; t^{m+1}_{i,j+1}(z,w) \\
&= -w\; t^{m+1}_{i+1,j}(z,w) + z\; t^{m+1}_{i+1,j+1}(z,w)
\end{aligned} \tag{3.7}
$$

for $0 \le i, j \le m$.

Proof. From (2.4), for $1 \le j \le m+1$, we obtain

$$
\begin{aligned}
&\sum_{i=0}^{m+1} t^{m+1}_{i,j}(z,w)\; z_1^{m+1-i} z_2^i \\
&\quad = (z z_1 - \overline{w} z_2)^{m+1-j} (w z_1 + \overline{z} z_2)^j \\
&\quad = \left\{ (z z_1 - \overline{w} z_2)^{m-(j-1)} (w z_1 + \overline{z} z_2)^{j-1} \right\} (w z_1 + \overline{z} z_2) \\
&\quad = \left\{ \sum_{i=0}^{m} t^m_{i,j-1}(z,w)\; z_1^{m-i} z_2^i \right\} (w z_1 + \overline{z} z_2) \\
&\quad = \sum_{i=0}^{m} t^m_{i,j-1}(z,w) \left\{ w z_1^{m+1-i} z_2^i + \overline{z} z_1^{m-i} z_2^{i+1} \right\} \\
&\quad = \sum_{i=0}^{m} t^m_{i,j-1}(z,w) w z_1^{m+1-i} z_2^i + \sum_{i=0}^{m} t^m_{i,j-1}(z,w) \overline{z} z_1^{(m+1)-(i+1)} z_2^{i+1} \\
&\quad = \sum_{i=0}^{m} w\; t^m_{i,j-1}(z,w) z_1^{m+1-i} z_2^i + \sum_{i=1}^{m+1} \overline{z}\; t^m_{i-1,j-1}(z,w) z_1^{m+1-i} z_2^i \\
&\quad = \sum_{i=0}^{m+1} \left\{ w\; t^m_{i,j-1}(z,w) + \overline{z}\; t^m_{i-1,j-1}(z,w) \right\} z_1^{m+1-i} z_2^i .
\end{aligned}
$$

Equating coefficients of $z_1^{m+1-i} z_2^i$ gives the second part of (3.6). The first part may be obtained in a similar manner, or more directly by application of the symmetry relation (3.2b) to both sides of the part just proved.

The backward recursion may be obtained directly from the forward recursion. Multiplying the first right-hand side of (3.6) by $\overline{z}$ and the second right-hand side (with j replaced by $j+1$) by $\overline{w}$ and summing gives the first right-hand side of (3.7). The second right-hand side is obtained similarly. ∎

These formulae are analogous to the addition formulae which express $\cos(m \pm 1)\theta$ and $\sin(m \pm 1)\theta$ in terms of $\cos\theta$, $\sin\theta$, $\sin m\theta$, and $\cos m\theta$.

LEMMA 3.4. *The outer functions* $o^m_{i,j}$ *defined by* (2.12) *satisfy the overlapping recurrence relations*

$$
o^{m+1}_{i,j}(z,w) = \begin{cases} \dfrac{\overline{z}}{|x|^2} o^m_{i,j}(z,w) + \dfrac{\overline{w}}{|x|^2} o^m_{i-1,j}(z,w), & 0 \le j \le m, \\ \dfrac{-w}{|x|^2} o^m_{i,j-1}(z,w) + \dfrac{z}{|x|^2} o^m_{i-1,j-1}(z,w), & 1 \le j \le m+1 \end{cases} \tag{3.8}
$$

for $0 \le i \le m+1$ *and also satisfy the backward recurrence relations*

$$
\begin{aligned}
o^m_{i,j}(z,w) &= z\; o^{m+1}_{i,j}(z,w) - \overline{w}\; o^{m+1}_{i,j+1}(z,w) \\
&= w\; o^{m+1}_{i+1,j}(z,w) + \overline{z}\; o^{m+1}_{i+1,j+1}(z,w)
\end{aligned} \tag{3.9}
$$

for $0 \le i, j \le m$.

Proof. This follows easily by substituting the expression (2.12) for $o^m_{i,j}$ into the relations (3.6) and (3.7) for the inner functions $t^m_{i,j}$. ∎

3.3. Derivatives and harmonicity. In this subsection we develop derivative formulae for the inner and outer functions. Applications include proofs of harmonicity, a dual basis for the inner functions, and an expression for the outer functions as appropriate derivatives of $1/|x|^2$. Analogous formulae in the three-dimensional case are given in Epton and Dembart [8].

We start by recalling the definitions for the complex derivative operators:

$$\begin{aligned} \frac{\partial}{\partial z} &= \frac{1}{2}\left(\frac{\partial}{\partial x_1} - \mathbf{i}\frac{\partial}{\partial x_2}\right), & \frac{\partial}{\partial \overline{z}} &= \frac{1}{2}\left(\frac{\partial}{\partial x_1} + \mathbf{i}\frac{\partial}{\partial x_2}\right), \\ \frac{\partial}{\partial w} &= \frac{1}{2}\left(\frac{\partial}{\partial x_3} - \mathbf{i}\frac{\partial}{\partial x_4}\right), & \frac{\partial}{\partial \overline{w}} &= \frac{1}{2}\left(\frac{\partial}{\partial x_3} + \mathbf{i}\frac{\partial}{\partial x_4}\right). \end{aligned} \tag{3.10}$$

By considering z, $\overline{z}$, w, and $\overline{w}$ as functions of x_1, x_2, x_3, and x_4, these operators may be applied to functions defined on $\mathbb{C}^2$ in the natural way. Indeed, immediate consequences of these definitions are the relations

$$\frac{\partial}{\partial z}\overline{z} = 0 \qquad \text{and} \qquad \frac{\partial}{\partial z} z = 1,$$

their conjugates and the similar relations with w, $\overline{w}$. More generally, if $f(x) = h(z, \overline{z}, w, \overline{w})$ is complex analytic in z, i.e., independent of $\overline{z}$ in the sense that $\partial h/\partial \overline{z} = 0$, then $\partial f/\partial z$ is given by all the usual rules for differentiation. Furthermore, armed with these operators the ordinary Laplacian may be expressed as

$$\Delta = 4\left(\frac{\partial^2}{\partial z \partial \overline{z}} + \frac{\partial^2}{\partial w \partial \overline{w}}\right). \tag{3.11}$$

LEMMA 3.5. *Any of the first partial derivatives map the inner functions of degree* m *to (multiples of) inner functions of degree* $m-1$. *Specifically, for* $0 \le i, j \le m+1$,

$$\frac{\partial}{\partial z} t^{m+1}_{i,j}(z,w) = (m+1-j)\, t^m_{i,j}(z,w), \tag{3.12a}$$

$$\frac{\partial}{\partial \overline{w}} t^{m+1}_{i,j}(z,w) = -(m+1-j)\, t^m_{i-1,j}(z,w), \tag{3.12b}$$

$$\frac{\partial}{\partial w} t^{m+1}_{i,j}(z,w) = j\, t^m_{i,j-1}(z,w), \tag{3.12c}$$

$$\frac{\partial}{\partial \overline{z}} t^{m+1}_{i,j}(z,w) = j\, t^m_{i-1,j-1}(z,w). \tag{3.12d}$$

Proof. Differentiating (2.4) with respect to z gives

$$\begin{aligned} \sum_{i=0}^{m+1} \frac{\partial}{\partial z} t^{m+1}_{i,j}(z,w) z_1^{m+1-i} z_2^i &= (m+1-j) z_1 (z_1 z - z_2 \overline{w})^{m-j} (z_1 w + z_2 \overline{z})^j \\ &= (m+1-j) z_1 \sum_{i=0}^{m} t^m_{i,j}(z,w) z_1^{m-i} z_2^i \\ &= (m+1-j) \sum_{i=0}^{m} t^m_{i,j}(z,w) z_1^{m+1-i} z_2^i. \end{aligned}$$

Equating coefficients gives (3.12a). In a similar manner, by differentiating (2.4) with respect to $\overline{w}$, w, and $\overline{z}$ we may obtain (3.12b), (3.12c), and (3.12d), respectively. ∎

LEMMA 3.6. *Any of the first partial derivatives map the outer functions of degree $-m-2$ to (multiples of) outer functions of degree $-m-3$. Specifically, for $m \geq 0$ and $0 \leq j \leq m$,*

$$\frac{\partial}{\partial z} o_{i,j}^m(z,w) = -(m+1-i)\; o_{i,j}^{m+1}(z,w), \qquad 0 \leq i \leq m+1, \tag{3.13a}$$

$$\frac{\partial}{\partial \overline{w}} o_{i,j}^m(z,w) = (m+1-i)\; o_{i,j+1}^{m+1}(z,w), \qquad 0 \leq i \leq m+1, \tag{3.13b}$$

$$\frac{\partial}{\partial w} o_{i,j}^m(z,w) = -(i+1)\; o_{i+1,j}^{m+1}(z,w), \qquad -1 \leq i \leq m, \tag{3.13c}$$

$$\frac{\partial}{\partial \overline{z}} o_{i,j}^m(z,w) = -(i+1)\; o_{i+1,j+1}^{m+1}(z,w), \qquad -1 \leq i \leq m. \tag{3.13d}$$

Proof of (3.13a). The proof is by induction on m.

Induction basis. Since

$$o_{0,0}^0(z,w) = (z\overline{z} + w\overline{w})^{-1},$$

an application of (3.8) shows

$$O_1(z,w) = |x|^{-4} \begin{bmatrix} \overline{z} & -w \\ \overline{w} & z \end{bmatrix}.$$

It follows that (3.13a) is true for $m = 0$.

Induction step. Assume that (3.13a) is true for $0 \leq m \leq M$. By the first part of (3.8),

$$o_{i,j}^{M+1}(z,w) = \frac{\overline{z}}{|x|^2} o_{i,j}^M(z,w) + \frac{\overline{w}}{|x|^2} o_{i-1,j}^M(z,w)$$

for $0 \leq i \leq M+1$ and $0 \leq j \leq M$. Using the inductive hypothesis to differentiate the outer functions,

$$\begin{aligned} \frac{\partial}{\partial z} o_{i,j}^{M+1}(x) =& \frac{\overline{z}(-\overline{z})}{|x|^4} o_{i,j}^M(x) + \frac{\overline{z}}{|x|^2} \frac{\partial}{\partial z} o_{i,j}^M(x) + \frac{\overline{w}(-\overline{z})}{|x|^4} o_{i-1,j}^M(x) + \frac{\overline{w}}{|x|^2} \frac{\partial}{\partial z} o_{i-1,j}^M(x) \\ =& \frac{-\overline{z}}{|x|^2} \left(\frac{\overline{z}}{|x|^2} o_{i,j}^M(x) + \frac{\overline{w}}{|x|^2} o_{i-1,j}^M(x) \right) - (M+1-i) \frac{\overline{z}}{|x|^2} o_{i,j}^{M+1}(x) \\ & - (M+2-i) \frac{\overline{w}}{|x|^2} o_{i-1,j}^{M+1}(x). \end{aligned}$$

Two applications of the first part of (3.8) give the result.

However, if $j = M+1$, the above does not hold. By the second part of (3.8),

$$o_{i,j}^{M+1}(z,w) = \frac{-w}{|x|^2} o_{i,j-1}^M(z,w) + \frac{z}{|x|^2} o_{i-1,j-1}^M(z,w)$$

for $0 \leq i \leq M+1$ and $1 \leq j \leq M+1$. Once again the inductive hypothesis will be used to differentiate this expression. This gives

$$\begin{aligned} \frac{\partial}{\partial z} o_{i,j}^{M+1}(x) =& \frac{(-w)(-\overline{z})}{|x|^4} o_{i,j-1}^M(x) + \frac{(-w)}{|x|^2} \frac{\partial}{\partial z} o_{i,j-1}^M(x) \\ & + \frac{z(-\overline{z})}{|x|^4} o_{i-1,j-1}^M(x) + \frac{z}{|x|^2} \frac{\partial}{\partial z} o_{i-1,j-1}^M(x) + \frac{1}{|x|^2} o_{i-1,j-1}^M(x) \end{aligned}$$

$$
\begin{aligned}
&= \frac{(-w)(-\overline{z})}{|x|^4} o^M_{i,j-1}(x) - (M+1-i)\frac{(-w)}{|x|^2} o^{M+1}_{i,j-1}(x) \\
&\quad + \frac{z(-\overline{z})}{|x|^4} o^M_{i-1,j-1}(x) - (M+2-i)\frac{z}{|x|^2} o^{M+1}_{i-1,j-1}(x) \\
&\quad + \frac{z\overline{z} + w\overline{w}}{|x|^4} o^M_{i-1,j-1}(x) \\
&= -(M+1-i)\frac{(-w)}{|x|^2} o^{M+1}_{i,j-1}(x) - (M+2-i)\frac{z}{|x|^2} o^{M+1}_{i-1,j-1}(x) \\
&\quad - \frac{(-w)}{|x|^2}\left(\frac{\overline{z}}{|x|^2} o^M_{i,j-1}(x) + \frac{\overline{w}}{|x|^2} o^M_{i-1,j-1}(x)\right).
\end{aligned}
$$

By the first part of (3.8), the term in the brackets is equal to $o^{M+1}_{i,j-1}(z,w)$ and thus the second part of (3.8) gives the required result. This proves (3.13a) by induction on m. The other three relations in (3.13) may be proven in a similar way. □

Given the above expressions for the derivatives of the inner and outer functions it is easy to show that they are harmonic functions. Since (3.11) may be used for the Laplacian, it follows that

(3.14)
$$
\frac{1}{4}\Delta o^m_{i,j} = \big(-(m+1-i)\big)\big(-(i+1)\big) o^{m+2}_{i+1,j+1} + (m+1-i)\big(-(i+1)\big) o^{m+2}_{i+1,j+1} = 0
$$

and

$$
\frac{1}{4}\Delta t^{m+2}_{i,j} = j(m+2-j)t^m_{i-1,j-1} - j(m+2-j)t^m_{i-1,j-1} = 0. \tag{3.15}
$$

Each of (3.14) and (3.15) can be inferred from the other as indicated in Remark 2.8.

From the product rule for the Laplacian,

$$
\Delta(fg) = (\Delta f)g + 2(\nabla f)\cdot(\nabla g) + f(\Delta g),
$$

and the Euler relation for a function f that is homogeneous of degree m,

$$
x\cdot\big(\nabla f\big)(x) = mf(x),
$$

we easily obtain the following.

LEMMA 3.7. *Let $|\cdot|$ be the 2-norm on $\mathbb{R}^d$, d being even. If $f:\mathbb{R}^d\setminus\{0\}\to\mathbb{R}$ is a nontrivial harmonic function that is homogeneous of degree m, then*

$$
\Delta(|\cdot|^{2\ell} f) = 4\ell\left(\frac{d}{2} + \ell + m - 1\right)|\cdot|^{2(\ell-1)} f.
$$

Hence $|\cdot|^{2\ell} f$ is polyharmonic of exact order

$$
\begin{cases} \ell+1 & \text{for } \ell \ge 0 \text{ and } m > -d/2 \text{ or } m < 1-\ell-d/2, \\ \ell+m+d/2 & \text{for } \ell < 0 \text{ and } m \ge 1-\ell-d/2. \end{cases}
$$

COROLLARY 3.8.

$$
\begin{aligned}
\Delta(|\cdot|^{2\ell} o^m_{i,j}) &= 4\ell(\ell-m-1)|\cdot|^{2(\ell-1)} o^m_{i,j}, \\
\Delta(|\cdot|^{2\ell} t^m_{i,j}) &= 4\ell(\ell+m+1)|\cdot|^{2(\ell-1)} t^m_{i,j}.
\end{aligned}
$$

All of the outer functions $o^m_{i,j}(x)$ may be written as (multiples of) derivatives of $1/|x|^2$ as in the following.

LEMMA 3.9. *Define*

$$R^m_{i,j} = \begin{cases} \dfrac{(-1)^{m-j+i}}{i!(m-i)!} \dfrac{\partial^m}{\partial \overline{w}^{j-i} \partial z^{m-j} \partial \overline{z}^i}, & i \le j, \\ \dfrac{(-1)^m}{i!(m-i)!} \dfrac{\partial^m}{\partial w^{i-j} \partial z^{m-i} \partial \overline{z}^j}, & i \ge j, \end{cases} \tag{3.16}$$

$0 \le i, j \le m$, $m \in \mathbb{N}_0$. *Then for* $x \in \mathbb{R}^4 \backslash \{0\}$,

$$o^m_{i,j}(x) = R^m_{i,j} \frac{1}{|x|^2}. \tag{3.17}$$

Furthermore, for harmonic functions f,

$$R^m_{i,j} R^{m'}_{i',j'} f = e(m, m', i)_{i',i'} R^{m+m'}_{i+i',j+j'} f, \tag{3.18}$$

where

$$e(m, m', i)_{i',i'} = \binom{i+i'}{i} \binom{m+m'-(i+i')}{m-i}. \tag{3.19}$$

Proof. The definition of the outer functions in terms of the inner functions, (2.12), and the fact that $t^0_{0,0} = 1$, imply

$$o^0_{0,0}(x) = \frac{1}{|x|^2}.$$

Then (3.17) follows by repeated use of Lemma 3.6.

Consider now (3.18). If $i \le j$ and $i' \le j'$ *or* $i \ge j$ and $i' \ge j'$, then (3.18) follows easily from (3.16). Turning to the slightly more difficult *mixed case*, assume, without loss of generality, $i \le j$ and $i' \ge j'$. Using definition (3.16),

$$R^m_{i,j} R^{m'}_{i',j'} f = \frac{(-1)^{m-j+i+m'}}{i!(m-i)!i'!(m'-i')!} \frac{\partial^{m+m'}}{\partial \overline{w}^{j-i} \partial w^{i'-j'} \partial z^{m-j+m'-i'} \partial \overline{z}^{i+j'}} f. \tag{3.20}$$

Let g be harmonic. Then from the form (3.11) for the Laplacian

$$\frac{\partial^2}{\partial z \partial \overline{z}} g = -\frac{\partial^2}{\partial w \partial \overline{w}} g.$$

Hence, if $i' - j' \ge j - i$, then

$$\frac{\partial^{i'-j'+j-i}}{\partial \overline{w}^{j-i} \partial w^{i'-j'}} f = (-1)^{j-i} \frac{\partial^{i'-j'+j-i}}{\partial w^{i+i'-(j+j')} \partial z^{j-i} \partial \overline{z}^{j-i}} f.$$

Substituting this into (3.20) gives

$$\begin{aligned} R^m_{i,j} R^{m'}_{i',j'} f &= \frac{(-1)^{m+m'}}{i!(m-i)!i'!(m'-i')!} \frac{\partial^{m+m'}}{\partial w^{i+i'-(j+j')} \partial z^{m+m'-(i+i')} \partial \overline{z}^{j+j'}} f \\ &= e(m, m', i)_{i',i'} R^{m+m'}_{i+i',j+j'} f, \end{aligned}$$

since $i' - j' \geq j - i$ implies $i + i' \geq j + j'$. The case when $i' - j' \leq j - i$ is similar. ∎

Motivated by the operators $R^m_{i,j}$ we are led to define operators

$$L^m_{i,j} = \begin{cases} \dfrac{1}{j!(m-j)!} \dfrac{\partial^m}{\partial w^{j-i} \partial \overline{z}^i \partial z^{m-j}}, & i \leq j, \\ \dfrac{(-1)^{i-j}}{j!(m-j)!} \dfrac{\partial^m}{\partial \overline{w}^{i-j} \partial \overline{z}^j \partial z^{m-i}}, & i \geq j. \end{cases} \tag{3.21}$$

Repeated use of Lemma 3.5 then shows the following.

LEMMA 3.10 (Dual basis for the inner functions). *The operators $L^m_{i,j}$ satisfy*

$$L^m_{i,j} t^m_{i,j} = t^0_{0,0},$$

and also have the more general property

$$L^{m'}_{i',j'} t^m_{i,j} = \binom{j}{j'} \binom{m-j}{m'-j'} t^{m-m'}_{i-i',j-j'}.$$

Thus the functionals

$$\lambda^m_{i,j}(f) = \big(L^m_{i,j} f\big)(0), \tag{3.22}$$

$0 \leq i, j \leq m$, $0 \leq m \leq p$, *form a dual basis for* $\big\{t^m_{i,j},\ 0 \leq i, j \leq m,\ 0 \leq m \leq p\big\}$.

4. Expansions of polyharmonic basic functions. In this section we develop far field expansions of the functions $|\cdot - x_<|^{-2}$ and $|\cdot - x_<|^{2n} \ln(|\cdot - x_<|)$, $n = 0, 1, \ldots$. These functions lead to polyharmonic splines of order 1 and $n + 2$, respectively. We also find bounds on the error in approximating the associated polyharmonic radial basis functions by truncating these far field expansions. We will truncate by dropping all terms of sufficiently negative homogeneity.

We will find it useful in this section to make use of the cosine formula in the form

$$\begin{aligned} |x - x_<|^2 &= |x|^2 + |x_<|^2 - 2\langle x, x_< \rangle \\ &= |x|^2 + |x_<|^2 - |x|^2 \operatorname{Tr}(x^{-1} x_<) \\ &= |x|^2 + |x_<|^2 - |x|^2 \chi_1(x^{-1} x_<) \qquad \text{(4.1a)} \\ &= |x|^2 + |x_<|^2 - \chi_1(x^* x_<) \\ &= |x|^2 + |x_<|^2 - \chi_1({x_<}^* x). \qquad \text{(4.1b)} \end{aligned}$$

4.1. Expansion of the potential. We start with a far field expansion of the potential function $|x - x_<|^{-2}$. This is an expansion in the character functions, which are products of powers of $|x_<|/|x|$ and the appropriate Gengenbauer polynomials—the Chebyshev polynomials of the second kind.

LEMMA 4.1. *For $x, x_< \in \mathbb{R}^4$ with $|x_<| < |x|$,*

$$\frac{1}{|x - x_<|^2} = \frac{1}{|x|^2} \sum_{m=0}^{\infty} \chi_m(x^{-1} x_<). \tag{4.2}$$

Proof. The result is trivially true when $x_< = 0$. Hence in what follows we assume that $0 < |x_<| < |x|$. Then from the definition of the character function, (2.9),

$$\chi_m(x^{-1} x_<) = \left(\frac{|x_<|}{|x|}\right)^m \frac{\sin(m+1)\theta}{\sin\theta},$$

where θ is the angle between x and $x_<$. As is well known,

$$\left|\frac{\sin(m+1)\theta}{\sin\theta}\right| \le m+1 \quad \text{for all} \quad \theta \in \mathbb{R}.$$

Therefore the series on the right of (4.2) converges absolutely.

We will prove the lemma by showing the product of the right-hand side of (4.2) with $|x-x_<|^2$ is 1. Let $y = x^{-1}x_<$, then

$$\begin{aligned}
|x-x_<|^2 \frac{1}{|x|^2}\sum_{m=0}^{\infty}\chi_m(y) &= \big\{|x|^2+|x_<|^2-|x|^2\chi_1(y)\big\}\frac{1}{|x|^2}\sum_{m=0}^{\infty}\chi_m(y)\\
&= \big\{1+|y|^2-\chi_1(y)\big\}\sum_{m=0}^{\infty}\chi_m(y)\\
&= \sum_{m=0}^{\infty}\big\{\chi_m(y)+|y|^2\chi_m(y)-\chi_1(y)\chi_m(y)\big\}\\
&= \sum_{m=0}^{\infty}\big\{\chi_m(y)+|y|^2\chi_m(y)-\chi_{m+1}(y)-|y|^2\chi_{m-1}(y)\big\}\\
&= \chi_0(y)-|y|^2\chi_{-1}(y)\\
&= 1,
\end{aligned}$$

where we have used Lemma 2.7 to expand the product $\chi_1(y)\chi_m(y)$, a telescoping argument which is valid since $\chi_m(y) \to 0$ as $m \to \infty$ and because of the fact that $\chi_{-1} = 0$. □

We will now obtain a bound on the error in approximating $\Phi_{x_<}(\cdot) = 1/|\cdot - x_<|^2$ by the truncated series

$$g_p(x) = \frac{1}{|x|^2}\sum_{m=0}^{p}\chi_m(x^{-1}x_<).$$

From the explicit formula for the character function, (2.9), $|\chi_m(y)| \le (m+1)|y|^m$ for all $y \in \mathbb{H}^1$ and therefore the error in approximating $\Phi_{x_<}(\cdot)$ by g_p is bounded by

$$\begin{aligned}
\big|\Phi_{x_<}(x)-g_p(x)\big| &\le \frac{1}{|x|^2}\sum_{m=p+1}^{\infty}\big|\chi_m(x^{-1}x_<)\big|\\
&\le \frac{1}{|x|^2}\sum_{m=p+1}^{\infty}(m+1)|y|^m\\
&= \frac{|y|^{p+1}}{|x|^2}\sum_{m=0}^{\infty}\big((p+1)+(m+1)\big)|y|^m\\
&= \frac{|y|^{p+1}}{|x|^2}\left\{(p+1)\sum_{m=0}^{\infty}|y|^m+\sum_{m=0}^{\infty}(m+1)|y|^m\right\},
\end{aligned}$$

where $y = x^{-1}x_<$. If $|x_<| < |x|$, then $|y| < 1$ and the well-known identities

$$\sum_{m=0}^{\infty}h^m = \frac{1}{1-h} \qquad \text{and} \qquad \sum_{m=0}^{\infty}(m+1)h^m = \frac{1}{(1-h)^2},$$

for $|h| < 1$, may be applied. This gives

$$\left|\Phi_{x_<}(x) - g_p(x)\right| \leq \frac{|y|^{p+1}}{|x|^2}\left\{\frac{p+1}{1-|y|} + \frac{1}{(1-|y|)^2}\right\}. \tag{4.3}$$

Denote the bound on the right of (4.3) by $e_p(|y|)$. For $|x|$ fixed, since each term on the right in (4.3) is obviously strictly increasing in $|y|$ for $0 < |y| < 1$, so is $e_p(|y|)$. Considering now the sum

$$s(x) = \sum_{k=1}^{N} \frac{d_k}{|x - x_k|^2},$$

we apply the bound above to each term and sum. The monotonicity of the bound enables us to estimate $e_p(|x_k|/|x|)$ by $e_p(d)$, where

$$d = \max_{1\leq k\leq N} \frac{|x_k|}{|x|}.$$

In combination with Lemma 4.1 and (2.14), this gives the following.

THEOREM 4.2. *Suppose $x_k \in \mathbb{R}^4$, $|x_k| \leq r$, and $d_k \in \mathbb{R}$ for each $1 \leq k \leq N$. Let*

$$s(x) = \sum_{k=1}^{N} \frac{d_k}{|x - x_k|^2},$$

and let C_m be the $(m+1)\times(m+1)$ matrix

$$\left[C^m_{i,j}\right] = \sum_{k=1}^{N} d_k\big(T_m(x_k)\big).$$

For $p \in \mathbb{N}_0$, set

$$g_p(x) = \sum_{m=0}^{p}\sum_{i,j=0}^{m} C^m_{j,i}\, o^m_{i,j}(x) = \sum_{m=0}^{p} \operatorname{Tr}\left(C_m O_m(x)\right), \tag{4.4}$$

$x \in \mathbb{R}^4\backslash\{0\}$. Then for all x with $|x| > r$

$$|s(x) - g_p(x)| \leq \frac{M}{r^2}\left(\frac{p+1}{1-1/c} + \frac{1}{(1-1/c)^2}\right)\left(\frac{1}{c}\right)^{p+3},$$

where $M = \sum_{k=1}^{N} |d_k|$ and $c = |x|/r$.

4.2. Expansion of a polyharmonic function. In this subsection our aim is to develop far field expansions for the functions $|\cdot - x_<|^{2n}\ln|\cdot - x_<|$, which are polyharmonic of order $n+2$. This will be done by induction on n with the biharmonic case $\ln|\cdot - x_<|$ being used as the induction basis.

The polyharmonic functions f of order $n+2$ that occur will be written in the form

$$f = f_0 + |\cdot|^2 f_1 + \cdots + |\cdot|^{2n+2} f_{n+1},$$

where $f_0, \ldots, f_{n+1}$ are harmonic. In this sum, $|\cdot|^{2j} f_j$ is actually a $(j+1)$-harmonic term. As a consequence the terms of a specified homogeneous order k in our expansions will no longer involve a single χ_m as in the harmonic case of Lemma 4.1. Rather, they will be a weighted sum of $\chi_m(y), |y|^2\chi_{m-2}(y), \ldots, |y|^{2n+2}\chi_{m-2n-2}(y)$, consistent with the polyharmonicity orders of Lemma 3.7. For this reason we need to know how pairs of character functions combine.

LEMMA 4.3. *For $m \in \mathbb{N}$, $m \geq 2$, and $|x_<| < |x|$,*

$$\chi_m(x^{-1}x_<) - \frac{|x_<|^2}{|x|^2}\chi_{m-2}(x^{-1}x_<) = 2\frac{|x_<|^m}{|x|^m}\cos(m\theta), \tag{4.5}$$

where θ is the angle between x and $x_<$.

Proof. The lemma is trivially true when $x_< = 0$. For $0 < |x_<| < |x|$, the explicit formula for the character function implies

$$\begin{aligned}
\chi_m(x^{-1}x_<) - \frac{|x_<|^2}{|x|^2}\chi_{m-2}(x^{-1}x_<) &= \frac{|x_<|^m}{|x|^m}\left(\frac{\sin(m+1)\theta}{\sin\theta} - \frac{\sin(m-1)\theta}{\sin\theta}\right) \\
&= \frac{|x_<|^m}{|x|^m}\left(\frac{\sin(m+1)\theta - \sin(m-1)\theta}{\sin\theta}\right) \\
&= 2\frac{|x_<|^m}{|x|^m}\frac{\sin(\theta)\cos(m\theta)}{\sin\theta} \\
&= 2\frac{|x_<|^m}{|x|^m}\cos(m\theta). \qquad \square
\end{aligned}$$

Our next goal is a far field expansion for the biharmonic function $\ln|\cdot - x_<|$. While this expansion is useful in and of itself, we will also use it as the induction basis for the expansion of the more general function $|\cdot - x_<|^{2n}\ln|\cdot - x_<|$, which is polyharmonic of order $n+2$.

LEMMA 4.4. *For $x, x_< \in \mathbb{R}^4$, and $|x_<| < |x|$,*

$$\ln|x - x_<|^2 = \ln|x|^2 - \sum_{m=1}^{\infty}\frac{1}{m}\left\{\chi_m(x^{-1}x_<) - \frac{|x_<|^2}{|x|^2}\chi_{m-2}(x^{-1}x_<)\right\}. \tag{4.6}$$

Proof. The case $x_< = 0$ is trivially true. Hence we assume $0 < |x_<| < |x|$.

We will use $\ln(\cdot)$ for the real logarithm and $\log(\cdot)$ for the principal branch of the complex logarithm. Thus $\Re(\log z) = \ln|z|$ away from the branch cut. We will represent $x = (x_1, x_2, x_3, x_4)^T \in \mathbb{R}^4$ by $x = [x_1 + \mathbf{i}x_2, x_3 + \mathbf{i}x_4] \in \mathbb{C}^2$ and similarly for $x_<$. There exists a rotation R_1 such that

$$R_1 x = [|x|, 0].$$

By the argument that precedes the introduction of the character function χ_m in (2.9), there is a rotation R_2 that fixes the north pole $[1,0]$ and rotates $R_1 x_<$ to the direction $[e^{\mathbf{i}\theta}, 0]$, where θ is the angle between $R_1 x_<$ and the north pole. Since rotations preserve angles θ is also the angle between x and $x_<$, $R = R_2 R_1$ is a rotation such that

$$Rx = [|x|, 0], \qquad Rx_< = [|x_<|e^{\mathbf{i}\theta}, 0].$$

Thus

$$|x - x_<| = |R(x - x_<)| = |Rx - Rx_<|$$
$$= \left|[|x| - |x_<|e^{\mathbf{i}\theta}, 0]\right| = |x|\left|1 - \frac{|x_<|}{|x|}e^{\mathbf{i}\theta}\right| = |x|\left|1 - |y|e^{\mathbf{i}\theta}\right|,$$

where $y = x^{-1}x_<$, and

$$\ln|x - x_<| = \Re\left\{\log\left(|x|\left(1 - |y|e^{\mathbf{i}\theta}\right)\right)\right\}$$
$$= \Re\left\{\log(|x|) + \log\left(1 - |y|e^{\mathbf{i}\theta}\right)\right\} = \ln|x| + \Re\left\{\log\left(1 - |y|e^{\mathbf{i}\theta}\right)\right\}.$$

Now

$$\log\left(1 - |y|e^{\mathbf{i}\theta}\right) = -\sum_{m=1}^{\infty}\frac{1}{m}\left(|y|e^{\mathbf{i}\theta}\right)^m = -\sum_{m=1}^{\infty}\frac{1}{m}|y|^m e^{\mathbf{i}m\theta}.$$

Taking the real part of this expression,

$$\Re\left\{\log\left(1 - \frac{|x_<|}{|x|}e^{\mathbf{i}\theta}\right)\right\} = -\sum_{m=1}^{\infty}\frac{1}{m}\frac{|x_<|^m}{|x|^m}\cos(m\theta)$$
$$= -\frac{1}{2}\sum_{m=1}^{\infty}\frac{1}{m}\left\{\chi_m(x^{-1}x_<) - \frac{|x_<|^2}{|x|^2}\chi_{m-2}(x^{-1}x_<)\right\},$$

where we have used Lemma 4.3 to express $(|x_<|^m/|x|^m)\cos(m\theta)$ in terms of the character functions χ_m. □

We now wish to obtain an expansion for the polyharmonic function

$$|x - x_<|^{2n}\ln|x - x_<|^2.$$

To simplify this procedure, we observe that

(4.7)
$$|x - x_<|^{2n}\ln|x - x_<|^2 = |x - x_<|^{2n}\ln|x|^2 + |x|^{2n}|I - x^{-1}x_<|^{2n}\ln|I - x^{-1}x_<|^2,$$

where I is the 2×2 identity in $\mathbb{H}_0^1$ or the element $[1, 0]$ in $\mathbb{C}^2$. This splits the function into a term containing the logarithmic singularity and a term amenable to "Laurent" expansion. We shall handle the two parts of the right-hand side of (4.7) separately.

The coefficient of $\ln|x|$ in the expansion. We first consider the polynomial that multiplies the $\ln|x|$ term in (4.7). We will give an expression for this polynomial in terms of the inner functions and discuss some symmetry properties.

LEMMA 4.5. *For $x, x_< \in \mathbb{R}^4$,*

$$|x - x_<|^{2n} = \sum_{m=0}^{n}|x|^{2m}\sum_{\ell=0}^{n-m}b_{m,\ell}^n|x_<|^{2\ell}\chi_{n-m-\ell}(x_<{}^*x) \tag{4.8}$$
$$= \sum_{m=0}^{n}|x|^{2m}\sum_{\ell=0}^{n-m}\sum_{i,j=0}^{n-m-\ell}D_{j,i}^{m,\ell}(x_<)t_{i,j}^{n-m-\ell}(x),$$

where the coefficients $b_{m,\ell}^n$ are given recursively by

$$b_{m,\ell}^{n+1} = b_{m-1,\ell}^n + b_{m,\ell-1}^n - b_{m,\ell}^n - b_{m-1,\ell-1}^n \tag{4.9a}$$

along with the initial conditions

$$b^0_{m,\ell} = \begin{cases} 1, & m = \ell = 0, \\ 0 & \text{otherwise}, \end{cases} \tag{4.9b}$$

$$b^n_{m,\ell} = 0 \qquad \text{if } m + \ell > n \text{ or } m, \ell \notin [0, n], \tag{4.9c}$$

and the coefficients $D^{m,\ell}_{j,i}(x_<)$ *are given by*

$$D^{m,\ell}_{j,i}(x_<) = b^n_{m,\ell} |x_<|^{2\ell} t^{n-m-\ell}_{j,i}(x_<{}^*). \tag{4.10}$$

Proof. Simple application of (4.1b), along with the product rule (2.10) for character functions χ_m, gives this first equality by induction on n. The second equality then follows by substituting (2.15) for $\chi_{n-m-\ell}(x_<{}^* x)$. ∎

Remark 4.6. Since $b^n_{m,\ell}$ is real and

$$t^{n-m-\ell}_{j,i}(x_<{}^*) = (-1)^{i+j} \overline{t^{n-m-\ell}_{n-m-\ell-j,n-m-\ell-i}(x_<{}^*)},$$

by symmetry (3.3) we see that

$$D^{m,\ell}_{j,i}(x_<) = (-1)^{i+j} \overline{D^{m,\ell}_{n-m-\ell-j,n-m-\ell-i}(x_<)}$$

for all $0 \le m \le n$, $0 \le \ell \le n - m$, $0 \le i, j \le n - m - \ell$. Provided the weights d_k are real, this symmetry is inherited by the coefficients of polynomials

$$q(x) = \sum_{k=0}^{N} d_k |x - x_k|^{2n} = \sum_{m=0}^{n} |x|^{2m} \sum_{\ell=0}^{n-m} \sum_{i,j=0}^{n-m-\ell} \widetilde{D}^{m,\ell}_{j,i} t^{n-m-\ell}_{i,j}(x) \tag{4.11}$$

occurring as the coefficient of the $\ln |x|^2$ in the truncated far field expansion of Theorem 4.10 to come.

One use of this property would be to recast the polynomial $q(x)$ as the weighted sum of approximately half as many $t^{n-m-\ell}_{i,j}(x)$'s, thereby reducing the operation count for approximate evaluation.

The nonlogarithmic part in the expansion. We now consider the infinite or far field part of (4.7). We find an explicit form for the expansion and give bounds on the error in approximation by truncation for this series.

LEMMA 4.7. *For* $n \in \mathbb{N}_0$ *and* $y \in \mathbb{R}^4$, $|y| < 1$,

$$|I - y|^{2n} \ln |I - y|^2 = \sum_{\ell=0}^{n+1} \sum_{m=\max\{1,2\ell\}}^{\infty} c^n_{m,\ell} |y|^{2\ell} \chi_{m-2\ell}(y), \tag{4.12}$$

where the series (4.12) *converges absolutely. The coefficients* $c^n_{m,\ell}$ *are given by the formulae*

$$c^0_{m,\ell} = \begin{cases} -1/m, & \ell = 0,\ m \ge 1, \\ 1/m, & \ell = 1,\ m - 2\ell \ge 0, \\ 0 & \text{otherwise} \end{cases} \tag{4.13a}$$

and the recurrence

$$c^{n+1}_{m,\ell} = c^n_{m,\ell} - c^n_{m-1,\ell} - c^n_{m-1,\ell-1} + c^n_{m-2,\ell-1}. \tag{4.13b}$$

Proof. The proof is by induction on n.

Induction basis: The case $n = 0$ of formulae (4.12) and (4.13a) is contained in Lemma 4.4.

Induction step: Assume (4.12) has been established for $n = K$. Then using the cosine formula (4.1a) and the product formula (2.10) we find

$$\begin{aligned}&|I-y|^{2K+2}\ln|I-y|^2\\&=\sum_{\ell=0}^{K+1}\sum_{m=\max\{1,2\ell\}}^{\infty}c_{m,\ell}^{K}|y|^{2\ell}\Big\{\chi_{m-2\ell}(y)-\chi_{m+1-2\ell}(y)-|y|^2\chi_{m-1-2\ell}(y)+|y|^2\chi_{m-2\ell}(y)\Big\}.\end{aligned}$$

Rearranging by collecting terms of the same homogeneity we find a series

$$|I-y|^{2K+2}\ln|I-y|^2=\sum_{\ell=0}^{K+2}\sum_{m=\max\{1,2\ell\}}^{\infty}c_{m,\ell}^{K+1}|y|^{2\ell}\chi_{m-2\ell}(y)$$

converging absolutely for $|y| < 1$ and with coefficients given by (4.13b). Thus (4.12) holds for $n = K + 1$. The result follows by induction. ∎

LEMMA 4.8. *For $m > 2n \geq 0$ and $0 \leq \ell \leq n+1$, the coefficients $c_{m,\ell}^{n}$ defined recursively in Lemma 4.7 have the explicit form*

$$c_{m,\ell}^{n}=(-1)^{n+\ell+1}n!\binom{n+1}{\ell}\Big/\prod_{\substack{k=m-\ell-n\\k\neq m-2\ell+1}}^{m-\ell+1}k. \tag{4.14}$$

Proof. The proof is by induction on n.

Induction basis: The formula for $n = 0$ is formula (4.13a) of Lemma 4.7.

Induction step: Assume that the formula holds for $n = K$ and $m > 2K$. Then for $n = K + 1$

$$c_{m,\ell}^{K+1}=c_{m,\ell}^{K}-c_{m-1,\ell}^{K}-c_{m-1,\ell-1}^{K}+c_{m-2,\ell-1}^{K}.$$

Assume now that $m > 2(K + 1)$ and $1 \leq \ell \leq K + 1$. Using the induction hypothesis,

$$\begin{aligned}c_{m,\ell}^{K+1}=&(-1)^{K+1+\ell}K!\Big\{\binom{K+1}{\ell}\Big\{\big(m-\ell-(K+1)\big)(m-2\ell+1)\\&-(m-\ell+1)(m-2\ell)\Big\}+\binom{K+1}{\ell-1}\Big\{(m-\ell-K-1)(m-2\ell+2)\\&-(m-\ell+1)(m-2\ell+1)\Big\}\Big\}\Big/\prod_{k=m-\ell-(K+1)}^{m-\ell+1}k\\=&(-1)^{K+1+\ell}\frac{K!(K+1)!}{\ell!(K+2-\ell)!}\Big\{(K+2-\ell)\Big\{(m-\ell-K-1)(m-2\ell+1)\\&-(m-\ell-1)(m-2\ell)\Big\}+\ell\Big\{(m-\ell-K-1)(m-2\ell+2)\\&-(m-\ell+1)(m-2\ell+1)\Big\}\Big\}\Big/\prod_{k=m-\ell-(K+1)}^{m-\ell+1}k\\=&(-1)^{K+2+\ell}\frac{K!(K+1)!}{\ell!(K+2-\ell)!}(K+2)(K+1)(m-2\ell+1)\Big/\prod_{k=m-\ell-(K+1)}^{m-\ell+1}k,\end{aligned}$$

agreeing with (4.14). The proof of the induction step when $\ell = 0$ or $\ell = K+2$ is similar. Hence the result follows by induction. ∎

LEMMA 4.9. *Let $n \in \mathbb{N}_0$ and $y \in \mathbb{R}^4$, $|y| < 1$. For $p \in \mathbb{N}$ let*

$$\widehat{g}_p(y) = \sum_{m=1}^{p} \sum_{\ell=0}^{\min\{\lfloor m/2 \rfloor, n+1\}} c_{m,\ell}^n |y|^{2\ell} \chi_{m-2\ell}(y). \tag{4.15}$$

If $p > 2n$, then

$$\left| |I-y|^{2n} \ln |I-y|^2 - \widehat{g}_p(y) \right| \le \frac{2^{n+1} n! (p+2)^2}{(p+1-n)\cdots(p-2n)} \frac{|y|^{p+1}}{1-|y|}.$$

Proof. By Lemma 4.7,

$$\left| |I-y|^{2n} \ln |I-y|^2 - \widehat{g}_p(y) \right| \le \left| \sum_{m=p+1}^{\infty} \sum_{\ell=0}^{n+1} c_{m,\ell}^n |y|^{2\ell} \chi_{m-2\ell}(y) \right|. \tag{4.16}$$

Then using Lemma 4.8 the magnitude of all the terms of order $|y|^m$ can be estimated by

$$\begin{aligned}
&\left| \sum_{\ell=0}^{n+1} c_{m,\ell}^n |y|^{2\ell} \chi_{m-2\ell}(y) \right| \\
&\qquad \le \sum_{\ell=0}^{n+1} \frac{(m-\ell-n-1)!}{(m-\ell+1)!} (m-2\ell+1) n! \binom{n+1}{\ell} (m-2\ell+1) |y|^m \\
&\qquad \le n! |y|^m \sum_{\ell=0}^{n+1} \frac{(m-2\ell+1)^2}{(m-\ell+1)\cdots(m-\ell-n)} \binom{n+1}{\ell}, \\
&\qquad \le n! |y|^m q_n(m) \sum_{\ell=0}^{n+1} \binom{n+1}{\ell} \\
&\qquad = n! 2^{n+1} |y|^m q_n(m),
\end{aligned}$$

where

$$q_n(m) = \frac{(m+1)^2}{\big(m-n\big)\big(m-(n+1)\big)\cdots\big(m-(2n+1)\big)}. \tag{4.17}$$

The function $q_n(m)$ is positive and decreasing in m for $m > 2n+1$. Hence the right-hand side of (4.16) can be estimated as

$$\begin{aligned}
\left| \sum_{m=p+1}^{\infty} \sum_{\ell=0}^{n+1} c_{m,\ell}^n |y|^{2\ell} \chi_{m-2\ell}(y) \right| &\le \sum_{m=p+1}^{\infty} \left| \sum_{\ell=0}^{n+1} c_{m,\ell}^n |y|^{2\ell} \chi_{m-2\ell}(y) \right| \\
&\le \sum_{m=p+1}^{\infty} n! 2^{n+1} |y|^m q_n(m) \\
&\le 2^{n+1} n! q_n(p+1) \frac{|y|^{p+1}}{1-|y|}. \qquad ∎
\end{aligned}$$

The full expansion and error bound. Combining the results from (4.7), Lemma 4.5, and Lemma 4.7, the function

$$\Phi_{x_<}(x) = |x - x_<|^{2n} \ln |x - x_<|^2$$

may be approximated by the truncated series

$$g_p(x) = \ln |x|^2 \sum_{m=0}^{n} |x|^{2m} \sum_{\ell=0}^{n-m} b_{m,\ell}^n |x_<|^{2\ell} \chi_{n-m-\ell}(x^* x_<) + |x|^{2n} \widehat{g}_p(x^{-1} x_<),$$

where $\widehat{g}_p$ is defined in (4.15). Then from Lemma 4.9 we obtain the error bound

$$\left|\Phi_{x_<}(x) - g_p(x)\right| \le |x|^{2n} \frac{2^{n+1} n! (p+2)^2}{(p+1-n)\cdots(p-2n)} \frac{|x^{-1}x_<|^{p+1}}{1 - |x^{-1}x_<|}$$

for $|x| > |x_<|$. Since $p > 2n$ this bound is increasing in $|y| = |x_<|/|x|$. We can apply the bound to each center x_k in turn and sum. This gives the following.

THEOREM 4.10. *Suppose $x_k \in \mathbb{R}^4$, $|x_k| \le r$, and $d_k \in \mathbb{R}$ for each $1 \le k \le N$. Let s be the $(n+2)$-harmonic spline*

$$s(x) = \sum_{k=1}^{N} d_k |x - x_k|^{2n} \ln |x - x_k|^2.$$

Furthermore, let $B_{m,\ell}$ be the $(n-m-\ell+1) \times (n-m-\ell+1)$ matrix

$$B_{m,\ell} = \left[B_{i,j}^{m,\ell}\right]_{i,j=0}^{n-m-\ell} = b_{m,\ell}^n \sum_{k=1}^{N} d_k |x_k|^{2\ell} \big(T_{n-m-\ell}(x_k^*)\big),$$

and let $C_{m,\ell}$ be the $(m-2\ell+1) \times (m-2\ell+1)$ matrix

$$C_{m,\ell} = \left[C_{i,j}^{m,\ell}\right]_{i,j=0}^{m-2\ell} = c_{m,\ell}^n \sum_{k=1}^{N} d_k |x_k|^{2\ell} \big(T_{m-2\ell}(x_k)\big),$$

where the coefficients $b_{m,\ell}^n$ and $c_{m,\ell}^n$ are given recursively by (4.9) *and* (4.13)*, respectively. Let $p \in \mathbb{N}$, $p > 2n$, and set*

(4.18)

$$\begin{aligned} g_p(x) &= \ln |x|^2 \sum_{m=0}^{n} \sum_{\ell=0}^{n-m} \sum_{i,j=0}^{n-m-\ell} B_{j,i}^{m,\ell} |x|^{2m} t_{i,j}^{n-m-\ell}(x) \\ &\qquad + \sum_{\ell=0}^{n+1} \sum_{m=\max\{1,2\ell\}}^{p} \sum_{i,j=0}^{m-2\ell} C_{j,i}^{m,\ell} |x|^{2(n+1-\ell)} o_{i,j}^{m-2\ell}(x) \\ &= \ln |x|^2 \sum_{m=0}^{n} |x|^{2m} \sum_{\ell=0}^{n-m} \operatorname{Tr}\big(B_{m,\ell} T_{n-m-\ell}(x)\big) \\ &\qquad + \sum_{\ell=0}^{n+1} \sum_{m=\max\{1,2\ell\}}^{p} |x|^{2(n+1-\ell)} \operatorname{Tr}\big(C_{m,\ell} O_{m-2\ell}(x)\big), \end{aligned}$$

$x \in \mathbb{R}^4 \backslash \{0\}$. Then for all x with $|x| > r$

$$\left|s(x) - g_p(x)\right| \le M r^{2n} \frac{(p+2)^2 2^{n+1} n!}{(p+1-n)\cdots(p-2n)} \left(\frac{1}{c}\right)^{p-2n+1} \frac{1}{1-1/c},$$

where $M = \sum_{k=1}^{N} |d_k|$ and $c = |x|/r$.

5. Uniqueness. In this section we will prove that the truncated expansions, g_p, appearing in (4.4) and (4.18) are the only functions of these forms achieving the stated asymptotic accuracy in approximating s as $|x| \to \infty$. These uniqueness results will allow us to form far field expansions in an inexpensive indirect manner, knowing that the expansions so obtained are identical with, and enjoy the same error estimates as, the computationally expensive directly formed expansions.

LEMMA 5.1. *Let $p \in \mathbb{N}_0$. Suppose a function $\widetilde{g}_p$ defined for $x \in \mathbb{R}^4 \backslash \{0\}$ can be written in the form*

$$\widetilde{g}_p(x) = \ln|x|^2 \sum_{m=0}^{n} \sum_{\ell=0}^{n-m} \sum_{i,j=0}^{n-m-\ell} \widetilde{B}_{j,i}^{m,\ell} |x|^{2m} t_{i,j}^{n-m-\ell}(x) + \sum_{\ell=0}^{n+1} \sum_{m=\max\{1,2\ell\}}^{p} \sum_{i,j=0}^{m-2\ell} \widetilde{C}_{j,i}^{m,\ell} |x|^{2(n+1-\ell)} o_{i,j}^{m-2\ell}(x), \tag{5.1}$$

where the various coefficients are complex numbers. Then

(i) *The coefficients $\{\widetilde{B}_{j,i}^{m,\ell}\}$ and $\{\widetilde{C}_{j,i}^{m,\ell}\}$ are uniquely determined by the function $\widetilde{g}_p$.*

(ii) *If $p > 2n$ and*

$$|\widetilde{g}_p(x)| = o(|x|^{2n-p}) \qquad \text{as } |x| \to \infty,$$

then $\widetilde{g}_p$ is identically zero.

Proof. We will need to use the fundamental properties of the inner and outer functions developed in section 3. Recall that $t_{i,j}^m$ is homogeneous of order m and $\{t_{i,j}^m : 0 \le i,j \le m,\ 0 \le m \le q\}$ is an orthogonal set of nontrivial spherical harmonics on the unit sphere $\mathcal{S}^3$. The definition of the outer functions (2.12)

$$o_{i,j}^m(z,w) = |x|^{-(2m+2)} t_{i,j}^m(\overline{z}, -w)$$

then implies that $o_{i,j}^m$ is homogeneous of order $-(m+2)$ and $\{o_{i,j}^m : 0 \le i,j \le m,\ 0 \le m \le q\}$ is linearly independent on $\mathcal{S}^3$.

Now fix $p \in \mathbb{N}_0$ and consider a function of the form (5.1). Rearrange the finite sum $\widetilde{g}_p$ by grouping together terms of the same growth at infinity, and arranging the groups in order of decreasing growth at infinity. The order of magnitude of $\widetilde{g}_p$ as $|x| \to \infty$ will be the same as that of the first nonzero group of terms.

Fix an integer k and denote the sum of the group of terms of growth $|x|^k \ln|x|$ by L_k. Thus

$$L_k(x) = \ln|x|^2 \sum_{m=0}^{n} \sum_{\ell=0}^{n-m} \sum_{i,j=0}^{n-m-\ell} \delta_{k,n+m-\ell} \widetilde{B}_{j,i}^{m,\ell} |x|^{2m} t_{i,j}^{n-m-\ell}(x).$$

Restricting attention to those terms for which the delta function is nonzero, we see that among these a particular inner function $t_{i',j'}^{k'}$ can arise only when $k' = k - 2m$ and thus can arise at most once. Hence by the linear independence of $\{t_{i,j}^m : 0 \le i,j \le m,\ 0 \le m \le n\}$ on $\mathcal{S}^3$, $L_k(x)$ is identically zero for all $x \ne 0$ if and only if

$$\delta_{k,n+m-\ell} \widetilde{B}_{j,i}^{m,\ell} = 0$$

for all $m = 0, \dots, n$; $\ell = 0, \dots, n-m$; $i, j = 0, \dots, n-m-\ell$. Similarly, the sum of the group of terms of growth $|x|^k$ at infinity is

$$G_k(x) = \sum_{\ell=0}^{n+1} \sum_{m=\max\{1,2\ell\}}^{p} \sum_{i,j=0}^{m-2\ell} \delta_{k,2n-m} \widetilde{C}_{j,i}^{m,\ell} |x|^{2(n+1-\ell)} o_{i,j}^{m-2\ell}.$$

Fix k and restrict attention to those terms for which the delta function is nonzero. A particular outer function $o_{i',j'}^{m'}$ can arise only when $m' = 2n-k-2\ell$. Since k and n are fixed this happens for at most one value of ℓ, and hence at most once. Thus by the linear independence of $\{o_{i',j'}^{m'} : 0 \le i', j' \le m',\ 0 \le m' \le q\}$ on $\mathcal{S}^3$, $G_k(x)$ is identically zero for all $x \neq 0$ if and only if

$$\delta_{k,2n-m} \widetilde{C}_{j,i}^{m,\ell} = 0$$

for all $\ell = 0, \dots, n+1$; $m = \max\{1, 2\ell\}, \dots, p$; $i, j = 0, \dots, m$.

For (ii) just note that if $p > 2n$, then $o(|x|^{2n-p}) \to 0$ so no terms can appear in the $\ln|x|^2$ summand, as each of those terms do not decay. Each term in the second summand is homogeneous of order $2n - m \ge 2n - p$. But the decay rate $o(|x|^{2n-p})$ precludes these terms from occurring, i.e., $\widetilde{g}_p = 0$. ∎

A simpler argument based on the same ideas shows the following.

LEMMA 5.2. *Let $p \in \mathbb{N}_0$. Suppose a function $\widetilde{g}_p$ defined for $x \in \mathbb{R}^4 \backslash \{0\}$ can be written in the form*

$$\widetilde{g}_p(x) = \sum_{m=0}^{p} \sum_{i,j=0}^{m} \widetilde{C}_{j,i}^{m}\, o_{i,j}^{m}(x).$$

Then

(i) *The coefficients $\{\widetilde{C}_{j,i}^{m}\}$ are uniquely determined by the function $\widetilde{g}_p$.*

(ii) *If*

$$|\widetilde{g}_p(x)| = o(|x|^{-(p+2)}) \qquad \text{as } |x| \to \infty,$$

then $\widetilde{g}_p$ is identically zero.

6. Translation of expansions. In this section we develop formulae which enable us to obtain a truncated expansion about one center indirectly from a truncated expansion about another. The operation count for this translation operation depends only on the order of the original expansion, not upon the number of centers x_k underlying it. In contrast, the operation count for direct expansion of a cluster is $\mathcal{O}(N(n+1)p^3)$, where N is the number of centers in the cluster. Thus indirect formation of expansions can be more efficient than direct expansion when the number of centers in a particular cluster is large and truncated expansions of subclusters are available.

For any matrix $A = (a_{i',j'})$, denote by $A|_{i,j}^{m}$ the $(m+1) \times (m+1)$ submatrix of A which begins at the (i, j) position, i.e.,

$$(A|_{i,j}^{m})_{i',j'} = a_{i+i',j+j'}, \qquad i', j' = 0, \dots, m.$$

THEOREM 6.1 (outer to outer or inner translation). *Let $x, x_< \in \mathbb{H}_0^1$, be such that $0 < |x_<| < |x|$. Then*

$$o_{i,j}^{m}(x - x_<) = \sum_{m'=0}^{\infty} \operatorname{Tr}\left(E(m, m', i) O_{m+m'}(x) \Big|_{i,j}^{m'} T_{m'}(x_<) \right), \tag{6.1}$$

where $E(m,m',i)$ *is the* $(m'+1)\times(m'+1)$ *diagonal matrix with entries*

$$e(m,m',i)_{i',i'} = \binom{i+i'}{i}\binom{m+m'-(i+i')}{m-i}.$$

Proof. Using Lemma 4.1, Lemma 2.9, and the relationship (3.17) between the outer functions and the operators $R^m_{i,j}$,

$$\begin{aligned}
o^m_{i,j}(x-x_<) = R^m_{i,j}\frac{1}{|x-x_<|^2} &= R^m_{i,j}\sum_{m'=0}^{\infty}\operatorname{Tr}\left(O_{m'}(x)T_{m'}(x_<)\right)\\
&= R^m_{i,j}\sum_{m'=0}^{\infty}\sum_{i',j'=0}^{m'} o^{m'}_{i',j'}(x)t^{m'}_{j',i'}(x_<)\\
&= R^m_{i,j}\sum_{m'=0}^{\infty}\sum_{i',j'=0}^{m'} t^{m'}_{j',i'}(x_<)R^{m'}_{i',j'}\frac{1}{|x|^2}\\
&= \sum_{m'=0}^{\infty}\sum_{i',j'=0}^{m'} t^{m'}_{j',i'}(x_<)R^m_{i,j}R^{m'}_{i',j'}\frac{1}{|x|^2}\\
&= \sum_{m'=0}^{\infty}\sum_{i',j'=0}^{m'} t^{m'}_{j',i'}(x_<)e(m,m',i)_{i',i'}R^{m+m'}_{i+i',j+j'}\frac{1}{|x|^2}\\
&= \sum_{m'=0}^{\infty}\sum_{i',j'=0}^{m'} e(m,m',i)_{i',i'}o^{m+m'}_{i+i',j+j'}(x)t^{m'}_{j',i'}(x_<)\\
&= \sum_{m'=0}^{\infty}\operatorname{Tr}\left(E(m,m',i)O_{m+m'}(x)\Big|^{m'}_{i,j}T_{m'}(x_<)\right),
\end{aligned}$$

where the differentiation term by term is justified by the real analyticity. ∎

This theorem is sufficient to translate a far field expansion of the type in (4.4). In particular consider an expansion like (4.4) but centered on $x_< \neq 0$. Then

$$\begin{aligned}
g_p(x-x_<) &= \sum_{m=0}^{p}\sum_{i,j=0}^{m} C^m_{j,i}\, o^m_{i,j}(x-x_<)\\
&= \sum_{m=0}^{p}\sum_{i,j=0}^{m} C^m_{j,i}\sum_{m'=0}^{\infty}\sum_{i',j'=0}^{m'} e(m,m',i)_{i',i'}o^{m+m'}_{i+i',j+j'}(x)t^{m'}_{j',i'}(x_<)\\
&= \sum_{m=0}^{p}\sum_{i,j=0}^{m} D^m_{j,i}\, o^m_{i,j}(x) + \mathcal{O}(|x|^{-(p+3)}),
\end{aligned} \tag{6.2}$$

where the coefficients $D^m_{j,i}$ are defined by the "convolution"

$$\binom{m}{i}D^m_{j,i} = \sum_{m'=0}^{m}\binom{m}{m'}\sum_{i',j'=0}^{m'}\binom{m-m'}{i-i'}t^{m-m'}_{j-j',i-i'}(x_<)\binom{m'}{i'}C^{m'}_{j',i'}. \tag{6.3}$$

Thus

$$h_p(x) = \sum_{m=0}^{p}\sum_{i,j=0}^{m} D^m_{j,i}\, o^m_{i,j}(x)$$

approximates $g_p(x)$ with error of order $\mathcal{O}(|x|^{-(p+3)})$ as $|x| \to \infty$. But by Theorem 4.2 the series formed directly, u_p, is of the same form and shares the same order of approximation. Hence the difference $u_p(x) - h_p(x)$ is $\mathcal{O}(|x|^{-(p+3)})$ as $|x| \to \infty$, and by the uniqueness theorem, Lemma 5.2, u_p and h_p are identical.

Furthermore, Theorem 6.1 is sufficient to translate an expansion of the form (6.2) into a Taylor series about 0 (a Maclaurin series). Since the functions $o^m_{i,j}$ are homogenous of degree m,

$$o^m_{i,j}(x - x_<) = (-1)^m o^m_{i,j}(x_< - x).$$

This simple observation allows the roles of x and $x_<$ to be switched in the application of Theorem 6.1. Starting with g_p defined by (6.2) and proceeding in this manner, we obtain

$$\begin{aligned} g_p(x - x_<) &= \sum_{m'=0}^{p} \sum_{i',j'=0}^{m'} C^{m'}_{j',i'} (-1)^{m'} o^{m'}_{i',j'}(x_< - x) \\ &= \sum_{m'=0}^{p} \sum_{i',j'=0}^{m'} (-1)^{m'} C^{m'}_{j',i'} \sum_{m=0}^{\infty} \sum_{i,j=0}^{m} e(m', m, i')_{i,i} o^{m+m'}_{i+i',j+j'}(x_<) t^m_{j,i}(x) \\ &= \sum_{m=0}^{p} \sum_{i,j=0}^{m} F^m_{j,i}\, t^m_{i,j}(x) + \mathcal{O}(|x|^{p+1}) \quad \text{as } |x| \to 0, \end{aligned}$$

where the coefficients $F^m_{j,i}$ are given by the "correlation"

(6.4)
$$\binom{m}{j}^{-1} F^m_{j,i} = \sum_{m'=0}^{p} \binom{m+m'}{m} \sum_{i',j'=0}^{m'} \binom{m+m'}{j+j'}^{-1} o^{m+m'}_{j+j',i+i'}(x_<)\, (-1)^{m'} \binom{m'}{j'} C^{m'}_{i',j'}.$$

Then by the characterization of the Maclaurin polynomial q of degree p for a function f as the only polynomial of total degree p with

$$|f(x) - q(x)| = o(|x|^p) \qquad \text{as } |x| \to 0,$$

it follows that

$$q(\cdot) = \sum_{m=0}^{p} \sum_{i,j=0}^{m} F^m_{j,i}\, t^m_{i,j}(\cdot). \tag{6.5}$$

THEOREM 6.2 (inner-to-inner translation formula). *For all $x, x_< \in \mathbb{R}^4$, $m \in \mathbb{N}_0$, and $0 \le i, j \le m$,*

$$t^m_{i,j}(x - x_<) = \sum_{m'=0}^{m} \sum_{\substack{i'=\max\{0,i-(m-m')\} \\ j'=\max\{0,j-(m-m')\}}}^{\substack{i'=\min\{i,m'\} \\ j'=\min\{j,m'\}}} (-1)^{m-m'} \binom{m-j}{m'-j'} \binom{j}{j'} t^{m-m'}_{i-i',j-j'}(x_<) t^{m'}_{i',j'}(x).$$

Proof. Because the functions $\left\{t^{m'}_{i',j'} : 0 \le i', j' \le m',\ 0 \le m' \le m\right\}$ form a basis for harmonic homogenous polynomials of degree at most m, and since $t^m_{i,j}(\cdot - x_<)$ is such a polynomial,

$$t^m_{i,j}(x - x_<) = \sum_{m'=0}^{m} \sum_{i',j'=0}^{m'} a^{m,m'}_{i,i',j,j'} t^{m'}_{i',j'}(x)$$

for some coefficients $a^{m,m'}_{i,i',j,j'}$ that depend on $x_<$. Applying the functionals $\lambda^{m''}_{i'',j''}$ of Lemma 3.10 to this expression gives

$$\binom{j}{j''}\binom{m-j}{m''-j''} t^{m-m''}_{i-i'',j-j''}(-x_<) = a^{m,m''}_{i,i'',j,j''}.$$

Since $(-1)^{m-m''}$ factors out by the homogeneity of $t^{m-m''}_{i-i'',j-j''}$, the result follows once we recall that the left-hand side is zero unless $j'' \le j$, $j - j'' \le m - m''$, $i'' \le i$, and $i - i'' \le m - m''$. ∎

This theorem may be used to translate a polynomial expansion such as (6.5). For example, if

$$q(x) = \sum_{m'=0}^{p} \sum_{i',j'=0}^{m'} F^{m'}_{j',i'}\, t^{m'}_{i',j'}(x - x_<), \tag{6.6}$$

then by Theorem 6.2 we get

$$\begin{aligned} q(x) &= \sum_{m'=0}^{p} \sum_{i',j'=0}^{m'} F^{m'}_{j',i'} \sum_{m=0}^{m'} \sum_{i,j=0}^{m} (-1)^{m'-m} \binom{m'-j'}{m-j} \binom{j'}{j} t^{m'-m}_{i'-i,j'-j}(x_<) t^{m}_{i,j}(x) \\ &= \sum_{m=0}^{p} \sum_{i,j=0}^{m} G^{m}_{j,i}\, t^{m}_{i,j}(x), \end{aligned}$$

where the coefficients $G^m_{j,i}$ are given by the "convolution"

$$\binom{m}{j}^{-1} G^m_{j,i} = \sum_{m'=0}^{p} \binom{m'}{m} \sum_{i',j'=0}^{m'} \binom{m'}{j'}^{-1} F^{m'}_{j',i'}\, (-1)^{m'-m} \binom{m'-m}{j'-j} t^{m'-m}_{i'-i,j'-j}(x_<). \tag{6.7}$$

It should be noted that this is an exact recentering of the polynomial q.

Just as we were able to translate expansions of the form (6.2), we want to be able to translate expansions like (4.18). One of our tools will be formulae for the products of z, w, $\overline{z}$, or $\overline{w}$ with a single inner or single outer function. These multiplication rules are contained in Lemmas 6.3 and 6.4 below.

LEMMA 6.3. *For $m \ge 0$ and $0 \le i, j \le m$,*

$$z\ o^m_{i,j}(z,w) = \frac{i+1}{m+1}\ |x|^2 o^{m+1}_{i+1,j+1}(z,w) + \frac{m-j}{m+1}\ o^{m-1}_{i,j}(z,w), \tag{6.8a}$$

$$w\ o^m_{i,j}(z,w) = -\frac{m+1-i}{m+1}\ |x|^2 o^{m+1}_{i,j+1}(z,w) + \frac{m-j}{m+1}\ o^{m-1}_{i-1,j}(z,w), \tag{6.8b}$$

$$\overline{z}\ o^m_{i,j}(z,w) = \frac{m+1-i}{m+1}\ |x|^2 o^{m+1}_{i,j}(z,w) + \frac{j}{m+1}\ o^{m-1}_{i-1,j-1}(z,w), \tag{6.8c}$$

$$\overline{w}\ o^m_{i,j}(z,w) = \frac{i+1}{m+1}\ |x|^2 o^{m+1}_{i+1,j}(z,w) - \frac{j}{m+1}\ o^{m-1}_{i,j-1}(z,w). \tag{6.8d}$$

Proof. First assume $m > 0$. Differentiate (2.13) with respect to $\overline{z}$. For the

right-hand side we obtain

$$\begin{aligned}(m-j)z_1(z_1\overline{z}+z_2\overline{w})^{m-1-j}(z_1(-w)+z_2z)^j&\\ &= (m-j)z_1|x|^{2m}\sum_{i=0}^{m-1} z_1^{m-1-i}z_2^i o_{i,j}^{m-1}(z,w)\\ &= (m-j)|x|^{2m}\sum_{i=0}^{m-1} z_1^{m-i}z_2^i o_{i,j}^{m-1}(z,w). \qquad (6.9)\end{aligned}$$

Since $|x|^2 = z\overline{z}+w\overline{w}$, for the left-hand side we have

$$\begin{aligned}&z(m+1)|x|^{2m}\sum_{i=0}^{m} z_1^{m-i}z_2^i o_{i,j}^m(z,w) + |x|^{2(m+1)}\sum_{i=0}^{m} z_1^{m-i}z_2^i\frac{\partial}{\partial\overline{z}}o_{i,j}^m(z,w) \qquad (6.10)\\ &= z(m+1)|x|^{2m}\sum_{i=0}^{m} z_1^{m-i}z_2^i o_{i,j}^m(z,w) - |x|^{2(m+1)}\sum_{i=0}^{m} z_1^{m-i}z_2^i(i+1)o_{i+1,j+1}^{m+1}(z,w),\end{aligned}$$

where we have used (3.13d) to evaluate $\frac{\partial}{\partial\overline{z}}o_{i,j}^m$. By considering the coefficient of $z_1^{m-i}z_2^i$ in (6.9) and (6.10) we see that

$$(m-j)o_{i,j}^{m-1}(z,w) = z(m+1)o_{i,j}^m(z,w) - |x|^2(i+1)o_{i+1,j+1}^{m+1}(z,w)$$

for $0\le i,j\le m$, which proves (6.8a)

By differentiating (2.13) with respect to $\overline{w}$, z, and w, in a similar manner we obtain (6.8b), (6.8c), and (6.8d), respectively, for $m>0$. The special case of $m=0$ for (6.8) follows directly from $o_{0,0}^0 = |x|^{-2}$ and the recurrence relations (3.8). □

Substituting (2.12) into (6.8) leads to a similar result for the inner functions. Specifically, we have the following.

LEMMA 6.4. *For $m\ge 0$ and $0\le i,j\le m$,*

$$z\ t_{i,j}^m(z,w) = \frac{m+1-i}{m+1}\ t_{i,j}^{m+1}(z,w) + \frac{j}{m+1}\ |x|^2 t_{i-1,j-1}^{m-1}(z,w), \qquad (6.11a)$$

$$w\ t_{i,j}^m(z,w) = \frac{m+1-i}{m+1}\ t_{i,j+1}^{m+1}(z,w) - \frac{m-j}{m+1}\ |x|^2 t_{i-1,j}^{m-1}(z,w), \qquad (6.11b)$$

$$\overline{z}\ t_{i,j}^m(z,w) = \frac{i+1}{m+1}\ t_{i+1,j+1}^{m+1}(z,w) + \frac{m-j}{m+1}\ |x|^2 t_{i,j}^{m-1}(z,w), \qquad (6.11c)$$

$$\overline{w}\ t_{i,j}^m(z,w) = -\frac{i+1}{m+1}\ t_{i+1,j}^{m+1}(z,w) + \frac{j}{m+1}\ |x|^2 t_{i,j-1}^{m-1}(z,w). \qquad (6.11d)$$

Since $t_{i,j}^m = 0$ and $o_{i,j}^m = 0$ if $m<0$, multiple applications of Lemma 6.3 and Lemma 6.4 may be used to obtain the following.

COROLLARY 6.5. *Let p be a given homogenous polynomial of degree m' in z, $\overline{z}$, w, and $\overline{w}$. Then to each inner function $t_{i,j}^m$ there correspond constants $\{F_{i',j'}^{\ell'}\}$ and to each outer function $o_{i,j}^m$ there correspond constants $\{G_{i',j'}^{\ell'}\}$ such that*

$$p(z,\overline{z},w,\overline{w})\ t_{i,j}^m(x) = \sum_{\ell'=0}^{\min\{m',\lfloor(m+m')/2\rfloor\}} |x|^{2\ell'} \sum_{i',j'=0}^{m+m'-2\ell'} F_{i',j'}^{\ell'} t_{i',j'}^{m+m'-2\ell'}(x),$$

$$p(z,\overline{z},w,\overline{w})\ o_{i,j}^m(x) = \sum_{\ell'=0}^{\min\{m',\lfloor(m+m')/2\rfloor\}} |x|^{2(m'-\ell')} \sum_{i',j'=0}^{m+m'-2\ell'} G_{i',j'}^{\ell'} o_{i',j'}^{m+m'-2\ell'}(x).$$

We now demonstrate how these results may be used to translate a truncated far field series, such as (4.18), due to a polyharmonic spline. Let g_p be such a series with center of expansion at $x_<$, i.e.,

$$g_p(x) = \ln|x-x_<|^2 \sum_{\ell=0}^{n} |x-x_<|^{2\ell} \left\{ \sum_{m=0}^{n-\ell} \sum_{i,j=0}^{n-\ell-m} B_{j,i}^{\ell,m} t_{i,j}^{n-\ell-m}(x-x_<) \right\} + \sum_{\ell=0}^{n+1} |x-x_<|^{2(n+1-\ell)} \left\{ \sum_{m=\max\{1,2\ell\}}^{p} \sum_{i,j=0}^{m-2\ell} C_{j,i}^{\ell,m} o_{i,j}^{m-2\ell}(x-x_<) \right\}. \tag{6.12}$$

The translations of objects such as those in the two sets of curly braces has been discussed already. The terms in the first set may be translated via Theorem 6.2 in much the same way as (6.6) was translated. The terms in the second set of braces may be translated using Theorem 6.1 in a similar manner to (6.2). Let $\{\widetilde{B}_{j,i}^{\ell,m}\}$ and $\{\widetilde{C}_{j,i}^{\ell,m}\}$ be the translated coefficients. Then

$$g_p(x) = \ln|x-x_<|^2 \sum_{\ell=0}^{n} |x-x_<|^{2\ell} \left\{ \sum_{m=0}^{n-\ell} \sum_{i,j=0}^{n-\ell-m} \widetilde{B}_{j,i}^{\ell,m} t_{i,j}^{n-\ell-m}(x) \right\} + \sum_{\ell=0}^{n+1} |x-x_<|^{2(n+1-\ell)} \left\{ \sum_{m=\max\{1,2\ell\}}^{p} \sum_{i,j=0}^{m-2\ell} \widetilde{C}_{j,i}^{\ell,m} o_{i,j}^{m-2\ell}(x) \right\} + \mathcal{O}(|x|^{-(p+1-2n)}). \tag{6.13}$$

Recall that

$$|x-x_<|^2 = |x|^2 - 2\langle x, x_<\rangle + |x_<|^2 = |x|^2 - (z\overline{z_<} + \overline{z}z_< + w\overline{w_<} + \overline{w}w_<) + |x_<|^2.$$

Thus we may use Lemma 6.4 to "translate" any product of the form $|x-x_<|^2 t_{i',j'}^{m'}(x)$ into a sum of at most ten terms of the form $|x|^{2\ell''} t_{i'',j''}^{m''}(x)$, where the coefficients of those terms depend on $x_<$. An analogous procedure employing Lemma 6.3 translates a product of the form $|x-x_<|^2 o_{i',j'}^{m'}(x)$ into a sum of at most ten terms of the form $|x|^{2\ell''} o_{i'',j''}^{m''}(x)$. Applying this procedure repeatedly to (6.13) viewed as a nested product

$$f_0(x) + |x-x_<|^2\Big(f_1(x) + |x-x_<|^2\Big(f_2(x) + \cdots + |x-x_<|^2\Big(f_n(x) + |x-x_<|^2 f_{n+1}(x)\Big)\cdots\Big)\Big)$$

brings g_p to the form

$$g_p(x) = \ln|x-x_<|^2 \sum_{\ell=0}^{n} |x|^{2\ell} \sum_{m=0}^{n-\ell} \sum_{i,j=0}^{n-\ell-m} \widetilde{\widetilde{B}}_{j,i}^{\ell,m}\ t_{i,j}^{n-\ell-m}(x) + \sum_{\ell=0}^{n+1} |x|^{2(n+1-\ell)} \sum_{m=\max\{1,2\ell\}}^{p} \sum_{i,j=0}^{m-2\ell} \widetilde{\widetilde{C}}_{j,i}^{\ell,m}\ o_{i,j}^{m-2\ell}(x) + \mathcal{O}(|x|^{-(p+1-2n)}) \tag{6.14}$$

with only the $\ln|x-x_<|^2$ term left untranslated. This step in the translation costs $\mathcal{O}((n+1)^2(p+1)^3)$ operations. This is acceptable since n is small, typically $n \le 2$, and the number of terms in the series to be translated is $\mathcal{O}((n+1)(p+1)^3)$.

By Theorem 4.10,

$$\ln|x-x_<|^2 = \ln|x|^2 + \sum_{\ell'=0}^{1} |x|^{2-2\ell'} \sum_{m'=\max\{1,2\ell'\}}^{p'} \sum_{i',j'=0}^{m'-2\ell'} D^{\ell',m'}_{j',i'}(x_<) o^{m'-2\ell'}_{i',j'}(x) + \mathcal{O}(|x|^{-(p'+1)}).$$

Substituting this into (6.14) gives

(6.15)
$$g_p(x) = \ln|x|^2 \sum_{\ell=0}^{n} |x|^{2\ell} \sum_{m=0}^{n-\ell} \sum_{i,j=0}^{n-\ell-m} \widetilde{\widetilde{B}}^{\ell,m}_{j,i}\ t^{n-\ell-m}_{i,j}(x) + F_{x_<}(x) + \sum_{\ell=0}^{n+1} |x|^{2(n+1-\ell)} \sum_{m=\max\{1,2\ell\}}^{p} \sum_{i,j=0}^{m-2\ell} \widetilde{\widetilde{C}}^{\ell,m}_{j,i}\ o^{m-2\ell}_{i,j}(x) + \mathcal{O}(|x|^{-(p+1-2n)}),$$

where $F_{x_<}(x)$ is given by the product

(6.16)
$$\left\{ \sum_{\ell=0}^{n} |x|^{2\ell} \sum_{m=0}^{n-\ell} \sum_{i,j=0}^{n-\ell-m} \widetilde{\widetilde{B}}^{\ell,m}_{j,i}\ t^{n-\ell-m}_{i,j}(x) \right\} \times \left\{ \sum_{\ell'=0}^{1} |x|^{2-2\ell'} \sum_{m'=\max\{1,2\ell'\}}^{p'} \sum_{i',j'=0}^{m'-2\ell'} D^{\ell',m'}_{j',i'}(x_<) o^{m'-2\ell'}_{i',j'}(x) \right\}$$

after it has been truncated by removing terms that are $\mathcal{O}(|x|^{-(p+1-2n)})$ as $|x| \to \infty$. From Corollary 6.5, (6.16) can be written in the form

$$\sum_{\ell=0}^{n+1} |x|^{2(n+1-\ell)} \sum_{m=\max\{1,2\ell\}}^{p} \sum_{i,j=0}^{m-2\ell} \widetilde{D}^{\ell,m}_{j,i} o^{m-2\ell}_{i,j}(x).$$

This completes the translation.

Appendix. A table of inner and outer functions.

$$T_0 = \begin{bmatrix} 1 \end{bmatrix},$$

$$T_1 = \begin{bmatrix} z & w \\ -\overline{w} & \overline{z} \end{bmatrix},$$

$$T_2 = \begin{bmatrix} z^2 & zw & w^2 \\ -2\,z\overline{w} & z\overline{z} - \overline{w}w & 2\,w\overline{z} \\ \overline{w}^2 & -\overline{w}\overline{z} & \overline{z}^2 \end{bmatrix},$$

$$T_3 = \begin{bmatrix} z^3 & z^2w & zw^2 & w^3 \\ -3\,z^2\overline{w} & z^2\overline{z} - 2\,\overline{w}zw & 2\,zw\overline{z} - \overline{w}w^2 & 3\,w^2\overline{z} \\ 3\,z\overline{w}^2 & -2\,z\overline{w}\overline{z} + \overline{w}^2w & z\overline{z}^2 - 2\,\overline{w}w\overline{z} & 3\,w\overline{z}^2 \\ -\overline{w}^3 & \overline{w}^2\overline{z} & -\overline{w}\overline{z}^2 & \overline{z}^3 \end{bmatrix},$$

$$T_4 = \begin{bmatrix} z^4 & z^3 w & z^2 w^2 & z w^3 & w^4 \\ -4\,z^3\overline{w} & z^3\overline{z} - 3\,\overline{w} z^2 w & 2\,z^2 w\overline{z} - 2\,\overline{w} z w^2 & 3\,z w^2\overline{z} - \overline{w} w^3 & 4\,w^3\overline{z} \\ 6\,z^2\overline{w}^2 & -3\,z^2\overline{w z} + 3\,z\overline{w}^2 w & z^2\overline{z}^2 - 4\,z\overline{w} w\overline{z} + \overline{w}^2 w^2 & 3\,z w\overline{z}^2 - 3\,\overline{w} w^2\overline{z} & 6\,w^2\overline{z}^2 \\ -4\,z\overline{w}^3 & 3\,z\overline{w}^2\overline{z} - \overline{w}^3 w & -2\,z\overline{w z}^2 + 2\,\overline{w}^2 w\overline{z} & z\overline{z}^3 - 3\,\overline{w} w\overline{z}^2 & 4\,w\overline{z}^3 \\ \overline{w}^4 & -\overline{w}^3\overline{z} & \overline{w}^2\overline{z}^2 & -\overline{w z}^3 & \overline{z}^4 \end{bmatrix},$$

$$O_0 = \begin{bmatrix} \dfrac{1}{z\overline{z} + w\overline{w}} \end{bmatrix},$$

$$O_1 = \frac{1}{|x|^4} \begin{bmatrix} \overline{z} & -w \\ \overline{w} & z \end{bmatrix},$$

$$O_2 = \frac{1}{|x|^6} \begin{bmatrix} \overline{z}^2 & -\overline{z} w & w^2 \\ 2\,\overline{z w} & \overline{z} z - \overline{w} w & -2\,w z \\ \overline{w}^2 & \overline{w} z & z^2 \end{bmatrix},$$

$$O_3 = \frac{1}{|x|^8} \begin{bmatrix} \overline{z}^3 & -\overline{z}^2 w & \overline{z} w^2 & -w^3 \\ 3\,\overline{z}^2\overline{w} & z\overline{z}^2 - 2\,\overline{w z} w & -2\,\overline{z} w z + \overline{w} w^2 & 3\,w^2 z \\ 3\,\overline{z w}^2 & 2\,\overline{z w} z - \overline{w}^2 w & \overline{z} z^2 - 2\,\overline{w} w z & -3\,w z^2 \\ \overline{w}^3 & \overline{w}^2 z & \overline{w} z^2 & z^3 \end{bmatrix},$$

$$O_4 = \frac{1}{|x|^{10}} \begin{bmatrix} \overline{z}^4 & -\overline{z}^3 w & \overline{z}^2 w^2 & -\overline{z} w^3 & w^4 \\ 4\,\overline{z}^3\overline{w} & z\overline{z}^3 - 3\,\overline{w z}^2 w & -2\,z\overline{z}^2 w + 2\,\overline{w z} w^2 & 3\,\overline{z} w^2 z - \overline{w} w^3 & -4\,w^3 z \\ 6\,\overline{z}^2\overline{w}^2 & 3\,z\overline{z}^2\overline{w} - 3\,\overline{z w}^2 w & \overline{z}^2 z^2 - 4\,\overline{z w} w z + \overline{w}^2 w^2 & -3\,\overline{z} w z^2 + 3\,\overline{w} w^2 z & 6\,w^2 z^2 \\ 4\,\overline{z w}^3 & 3\,\overline{z w}^2 z - \overline{w}^3 w & 2\,\overline{z w} z^2 - 2\,\overline{w}^2 w z & \overline{z} z^3 - 3\,\overline{w} w z^2 & -4\,w z^3 \\ \overline{w}^4 & \overline{w}^3 z & \overline{w}^2 z^2 & \overline{w} z^3 & z^4 \end{bmatrix}.$$

$$|x|^2 = z\overline{z} + w\overline{w}$$

REFERENCES

[1] J. F. Adams, *Lectures on Lie Groups*, W. A. Benjamin, Inc., New York, 1969.

[2] S. L. Altmann, *Rotations, Quaternions and Double Groups*, Oxford University Press, New York, 1986.

[3] R. K. Beatson and L. L. Greengard, *A short course on fast multipole methods*, in Wavelets, Multilevel Methods and Elliptic PDEs, M. Ainsworth, J. Levesley, W. Light, and M. Marletta, eds., Oxford University Press, New York, 1997, pp. 1–37.

[4] R. K. Beatson and W. A. Light, *Fast evaluation of radial basis functions: Methods for two-dimensional polyharmonic splines*, IMA J. Numer. Anal., 17 (1997), pp. 343–372.

[5] R. K. Beatson and G. N. Newsam, *Fast evaluation of radial basis functions:* I, Comput. Math. Appl., 24 (1992), pp. 7–19.

[6] R. K. Beatson and G. N. Newsam, *Fast evaluation of radial basis functions: Moment-based methods*, SIAM J. Sci. Comput., 19 (1998), pp. 1428–1449.

[7] R. K. Beatson, A. M. Tan, and M. J. D. Powell, *Fast Evaluation of Radial Basis Functions: Methods for* 3*-Dimensional Polyharmonic Splines*, in preparation.

[8] M. A. Epton and B. Dembart, *Multipole translation theory for the three-dimensional Laplace and Helmholtz equations*, SIAM J. Sci. Comput., 16 (1995), pp. 865–897.

[9] J. Flusser, *An adaptive method for image registration*, Pattern Recognition, 25 (1992), pp. 45–54.

[10] R. Franke, *Scattered data interpolation: Tests of some methods*, Math. Comp., 38 (1982), pp. 181–200.

[11] L. L. Greengard and V. Rokhlin, *A fast algorithm for particle simulations*, J. Comput. Phys, 73 (1987), pp. 325–348.
[12] E. Hewitt and K. A. Ross, *Abstract Harmonic Analysis,* II, Grundlehren Math. Wiss. 152, Springer-Verlag, Berlin, 1970.
[13] M. F. Hutchinson, *The application of thin plate smoothing splines to continent-wide data assimilation*, in Data Assimilation Systems, J. D. Jasper, ed., BRMC Research Report No. 27, Melbourne, 1991, Bureau of Meteorology, pp. 104–113.
[14] A. Kyrala., *Theoretical Physics: Applications of Vectors, Matrices, Tensors and Quaternions*, W. B. Saunders Company, Philadelphia, 1967.
[15] C. Müller, *Spherical Harmonics*, Lecture Notes in Math. 17, Springer-Verlag, New York, 1966.
[16] M. A. Naimark and A. I. Stern, *Theory of Group Representations*, Grundlehren Math. Wiss. 246, Springer-Verlag, New York, 1982.
[17] M. J. D. Powell, *The theory of radial basis function approximation in* 1990, in Advances in Numerical Analysis II: Wavelets, Subdivision Algorithms and Radial Functions, W. Light, ed., Oxford University Press, New York, 1992, pp. 105–210.
[18] D. L. Ragozin, *Constructive polynomial approximation on spheres and projective spaces*, Trans. Amer. Math. Soc., 162 (1971), pp. 157–170.
[19] D. L. Ragozin, *Uniform convergence of spherical harmonic expansions*, Math. Ann., 195 (1972), pp. 87–94.
[20] R. Sibson and G. Stone, *Computation of thin-plate splines*, SIAM J. Sci. Statist. Comput., 12 (1991), pp. 1304–1313.
[21] G. Turk and J. F. O'Brien, *Shape transformation using variational implicit surfaces*, in SIGGRAPH'99, Los Angeles, ACM, New York, 1999, pp. 335–342.

SIAM J. MATH. ANAL.
Vol. 32, No. 6, pp. 1311–1323

SPACE-TIME PERIODIC SOLUTIONS AND LONG-TIME BEHAVIOR OF SOLUTIONS TO QUASI-LINEAR PARABOLIC EQUATIONS*

G. BARLES† AND P. E. SOUGANIDIS‡

Abstract. This paper considers (i) the existence of space-time periodic solutions of quasi-linear parabolic equations and (ii) the convergence, as $t \to \infty$, of space periodic solutions of the initial value problem of quasi-linear parabolic equations to the space-time periodic solutions.

Key words. asymptotic behavior, space-time periodic solutions, quasi-linear parabolic equations, strong maximum principle, viscosity solutions

AMS subject classifications. 35K45, 35K65, 35B10, 35B27, 35B40, 35B60

PII. S0036141000369344

1. Introduction.

In this article, we are interested in the following two questions about space-time periodic quasi-linear parabolic equations: (i) the existence of space-time periodic solutions and (ii) the asymptotic behavior as $t \to \infty$ of the solutions of the initial value problem for quasi-linear parabolic equations in the periodic setting.

Concerning the first question, the typical result is that there exists a unique $\lambda \in \mathbb{R}$ such that the equation

$$\phi_t - \mathrm{tr}(A(x,t,D\phi)D^2\phi) + H(x,t,D\phi) + \lambda = 0 \quad \text{in } \mathbb{R}^N \times \mathbb{R} \tag{1.1}$$

has a space-time periodic solution $\phi : \mathbb{R}^N \times \mathbb{R} \to \mathbb{R}$ provided A and H are space-time periodic in (x,t). It is necessary, of course, to impose a number of conditions on A and H; these are stated later.

The second question is about the asymptotic behavior, as $t \to +\infty$, of the solution of the initial value problem

$$\begin{cases} u_t - \mathrm{tr}(A(x,t,Du)D^2u) + H(x,t,Du) = 0 & \text{in } \mathbb{R}^N \times (0,+\infty), \\ u(x,0) = u_0(x) & \text{in } \mathbb{R}^N \times \{0\} \end{cases} \tag{1.2}$$

with $u_0 \in W^{1,\infty}(\mathbb{R}^N)$ and periodic. The type of result expected here is that, as $t \to \infty$,

$$u(x,t) - \lambda t - \phi(x,t) \to 0 \quad \text{uniformly in } \mathbb{R}^N, \tag{1.3}$$

where ϕ is a space-time periodic solution of (1.1).

To explain the results we obtain here, we consider first the model case in which $A(x,t,p) \equiv$ Id, i.e., the case where (1.1) has the form

$$\phi_t - \Delta\phi + H(x,t,D\phi) + \lambda = 0 \quad \text{in } \mathbb{R}^N \times \mathbb{R} \tag{1.4}$$

*Received by the editors March 20, 2000; accepted for publication (in revised form) October 16, 2000; published electronically March 15, 2001.

http://www.siam.org/journals/sima/32-6/36934.html

†Laboratoire de Mathématiques et Physique Théorique, Faculté des Sciences et Techniques, Université de Tours, Parc de Grandmont, 37200 Tours, France (barles@univ-tours.fr). This author's work was partially supported by the TMR program "Viscosity Solutions and Their Applications."

‡Department of Mathematics, The University of Texas at Austin, Austin, TX 78712 (souganid@math.utexas.edu). This author's work was partially supported by the NSF and the TMR programs "Viscosity Solutions and Their Applications" and "Hyperbolic Conservation Laws."

with H a space-time periodic, continuous function.

A classical method for proving the existence of ϕ and λ consists in considering the problem

$$\phi_t^\epsilon - \Delta\phi^\epsilon + H(x,t,D\phi^\epsilon) + \epsilon\phi^\epsilon = 0 \quad \text{in } \mathbb{R}^N \times \mathbb{R}\,. \tag{1.5}$$

Under suitable assumptions on H, the existence of a space-time periodic solution of (1.5) follows easily from the maximum principle and the fixed point theorem for contraction maps. The problem then is to let $\epsilon \to 0$. Although the family $(\phi^\epsilon)_{\epsilon>0}$ does not converge in general, the goal is to show the existence of a constant c_ϵ such that $\phi^\epsilon(x,t) - c_\epsilon$ converges, locally uniformly, at least along a subsequence, and that ϵc_ϵ converges to the constant λ. In most cases it suffices to choose the constant $c_\epsilon = \phi^\epsilon(x_0,t_0)$ for some $(x_0,t_0) \in \mathbb{R}^N \times \mathbb{R}$, although occasionally other choices are more suitable.

In order to obtain this type of behavior for the family $(\phi^\epsilon)_{\epsilon>0}$, it is necessary to establish a gradient bound. In the case at hand, this will be a bound on $D\phi^\epsilon$ and not on ϕ_t^ϵ. This is where the properties of H play a key role. We consider two cases, which we call *sublinear* and *superlinear*, despite the fact that the terminology is not completely accurate.

In the *sublinear case* the gradient bound comes from the ellipticity of the equation, i.e., from the Δ-term in (1.4). To obtain this gradient bound, it is necessary to impose certain growth conditions on $H(x,t,p)$ with respect to p, which are, in fact, a bit more general than sublinearity. Our approach here is inspired by the method of Ishii and Lions [10] (see also Barles [2]).

On the contrary, in the *superlinear case* the gradient bound comes from the H-term. Therefore, our approach can handle cases of degenerate parabolic equations and even first-order Hamilton–Jacobi equations. The main tool here is the "weak Bernstein-type arguments" introduced in Barles [1].

It is worth mentioning that the main difficulty in obtaining gradient bounds is the fact that we cannot use any L^∞-bounds on the ϕ^ϵ. In fact the opposite is true. The gradient bounds will be used to deduce some kind of L^∞-bounds on ϕ^ϵ, namely, bounds on $\phi^\epsilon - c_\epsilon$. As a matter of fact, and this is perhaps a little bit surprising, the gradient bound in x is enough to get the full answer, i.e., the convergence of $\phi^\epsilon - c_\epsilon$ along subsequences. In particular, no bounds on ϕ_t^ϵ are needed.

The approach described above yields rather general existence results of space-time periodic solutions both for (1.4) and (1.1). Unfortunately our results on the asymptotic behavior as $t \to +\infty$ of the solutions of (1.2) are proved under far more restrictive assumptions. In fact, the key ingredient will be the strong maximum principle, which will require in the case of (1.1)–(1.2) the smoothness of solutions. It is therefore necessary to impose appropriate assumptions on A and H in order to be able to prove the needed regularity. In particular, we are able to treat only the uniformly parabolic case. It is worth remarking here that in the completely degenerate case, i.e., when $A = 0$, it is not true that solutions of (1.2) converge as $t \to \infty$ to solutions of (1.1), as was shown recently by Fathi and Mather [8] and Barles and Souganidis [3]. Of course, more classical results can be used for (1.4) and the study of the asymptotic behavior of the solution of the associated initial value problem.

The existence of space-time periodic solutions for parabolic equations has been studied in Namah and Roquejoffre [17] using different methods and, in particular, degree theory. The main assumption of [17] is that H in (1.2) depends on u, which implies that (1.3) holds with an exponential rate of convergence. Roquejoffre [18], [19]

also considered the case of first-order Hamilton–Jacobi equations with strictly convex nonlinearity and obtained asymptotic results using the approach of [16]. Finally, when $A \equiv 0$, the existence of λ and of a space-time periodic solution of (1.1) was also studied in Majda and Souganidis [15] as a step in the study of some models in turbulent combustion.

The existence of a space-time periodic solution of parabolic equations is also related to the ergodic properties of stochastic processes in space-time environments, for example, the existence of principal eigenvalues for uniformly parabolic space-time periodic operators. Such questions have been looked at in different contexts by a number of authors but not in the generality of this paper. We refer to Freidlin [9], Sinai [20], and the references therein.

The paper is organized as follows. In section 2 we consider the case of the model equation (1.4). Section 3 is devoted to the extensions to the general equation (1.1). In section 4 we provide the asymptotic results for (1.2).

2. The model equation.

2.1. The sublinear case.

The main assumption on H is

$$\text{(H1)} \qquad \begin{cases} H \in C(\mathbb{R}^N \times \mathbb{R} \times \mathbb{R}^N) \text{ is } [0,\Lambda]^N \times [0,T]\text{-space-time periodic, and} \\ \text{there exists a continuous function } \chi : [0,+\infty) \to \mathbb{R} \text{ such that} \\ \displaystyle\int^{+\infty} \chi(u)^{-1} du = +\infty \text{ and } |H(x,t,p)| \le \chi(|p|) \text{ in } \mathbb{R}^N \times \mathbb{R} \times \mathbb{R}^N. \end{cases}$$

THEOREM 2.1. *Assume* (H1). *Then there exists at least one solution* (ϕ, λ) *of* (1.4), *where* ϕ *is a continuous space-time periodic function and* $\lambda \in \mathbb{R}$. *The constant* λ *is unique. Moreover, if* H *is locally Lipschitz continuous in* p, *then* ϕ *is unique up to constants.*

Proof. 1. The uniqueness of λ, which is classical even in the context of viscosity solutions theory (see, for example, Lions, Papanicolaou, and Varadhan [14] or, in a slightly different context, Lions [13]), follows from the maximum principle. The up to constants uniqueness of ϕ is not true in general. Here it follows from the strong maximum principle.

2. The existence is based on a priori estimates on the gradient of the space-time periodic solution ϕ^ϵ of the equation

$$\phi_t^\epsilon - \Delta\phi^\epsilon + H(x,t,D\phi^\epsilon) + \epsilon\phi^\epsilon = 0 \quad \text{in } \mathbb{R}^N \times \mathbb{R}. \tag{2.1}$$

□

The key result is the following lemma.

LEMMA 2.2. *If* (H1) *holds and* ϕ^ϵ *is a continuous, space-time periodic solution of* (2.1), *then there exists* $K > 0$ *depending only on* H *such that*

$$\sup_{t \in \mathbb{R}} \|D\phi^\epsilon(\cdot,t)\|_\infty \le K.$$

Proof. 1. Consider the

$$\max_{\mathbb{R}^N \times \mathbb{R}^N \times \mathbb{R}} \left[\phi^\epsilon(x,t) - \phi^\epsilon(y,t) - \psi(|x-y|)\right]$$

for a suitable increasing concave function $\psi : [0,+\infty) \to \mathbb{R}$ such that $\psi(0) = 0$. This is indeed a maximum, since ϕ^ϵ is (x,t)-periodic and ψ is increasing. Hence the max is achieved in the set

$$Q_\Lambda = \{(x,y,t) : |x-y| \le 2\Lambda, t \in [0,T]\}.$$

It follows that ψ has to be built only on $[0, 2\Lambda]$, and then it suffices to extend it to $[0, +\infty)$ by keeping it increasing and concave. To this end, if χ is the function given by (H1), we choose ψ to be a solution of

$$2\psi'' = -\chi(\psi') \quad \text{on } (0, 2\Lambda), \quad \psi(0) = 0$$

with $\psi' > 0$ on $[0, 2\Lambda]$. Since ψ' can be defined by

$$2\int_{\psi'(s)}^{K} \chi(u)^{-1} du = s,$$

it follows from (H1) that, if K is chosen large enough, ψ satisfies all the above requirements.

2. If the max is zero—note that it is always nonnegative—then, for all x, y, t,

$$\phi^\epsilon(x,t) - \phi^\epsilon(y,t) \le \psi(|x-y|) \le K|x-y|,$$

the last inequality being a consequence of the concavity of ψ.

3. If the max is positive, let $(\overline{x}, \overline{y}, \overline{t})$ be the point where it is achieved. It follows that $\overline{x} \neq \overline{y}$. Then classical results from the theory of viscosity solutions (see the user's guide by Crandall, Ishii, and Lions [7]) yield the existence of $a \in \mathbb{R}$ and symmetric matrices X, Y such that $(a, p, X) \in D^{2,+}\phi^\epsilon(\overline{x}, \overline{t})$, $(a, p, Y) \in D^{2,-}\phi^\epsilon(\overline{y}, \overline{t})$ and

$$\begin{pmatrix} X & 0 \\ 0 & -Y \end{pmatrix} \le \psi'(|\overline{x}-\overline{y}|) \begin{pmatrix} B & -B \\ -B & B \end{pmatrix} + \psi''(\overline{x}-\overline{y}) \begin{pmatrix} q\otimes q & -q\otimes q \\ -q\otimes q & q\otimes q \end{pmatrix}, \tag{2.2}$$

where $p = |\overline{x}-\overline{y}|^{-1}\psi'(|\overline{x}-\overline{y}|)(\overline{x}-\overline{y})$, $q = |\overline{x}-\overline{y}|^{-1}(\overline{x}-\overline{y})$, and $B = |\overline{x}-\overline{y}|^{-1}(I - q\otimes q)$.

Applying (2.2) to any $\mathbb{R}^{2N}$-vector of the form (r, r) with $r \in \mathbb{R}^N$ gives

$$X \le Y.$$

To obtain the gradient bounds, we follow Ishii and Lions [10] and apply (2.2) to $(q, -q)$ using that $Bq = 0$. It follows that

$$(Xq, q) - (Yq, q) \le 4\psi''(|\overline{x}-\overline{y}|)\ .$$

Combining the above we obtain

$$\operatorname{tr}(X) - \operatorname{tr}(Y) \le 4\psi''(|\overline{x}-\overline{y}|)\ . \tag{2.3}$$

On the other hand, the viscosity inequalities for (2.1) read

$$a - \operatorname{tr}(X) + H(\overline{x}, \overline{t}, p) + \epsilon\phi^\epsilon(\overline{x}, \overline{t}) \le 0 \quad \text{and} \quad a - \operatorname{tr}(Y) + H(\overline{y}, \overline{t}, p) + \epsilon\phi^\epsilon(\overline{y}, \overline{t}) \ge 0.$$

Subtracting these last two inequalities and using (2.3) together with (H1) gives

$$-4\psi''(|\overline{x}-\overline{y}|) - 2\chi(\psi'(|\overline{x}-\overline{y}|)) + \epsilon[\phi^\epsilon(\overline{x}, \overline{t}) - \phi^\epsilon(\overline{y}, \overline{t})] \le 0.$$

Since $2\psi'' = -\chi(\psi')$ and $\phi^\epsilon(\overline{x}, \overline{t}) - \phi^\epsilon(\overline{y}, \overline{t}) > 0$ (otherwise the maximum would be negative), this is a contradiction. Hence the maximum is always zero, and the claim follows by step 2. □

Remark 2.3. The above proof does not use the L^∞-norm of ϕ^ϵ. The gradient bound is therefore independent of $\|\phi^\epsilon\|_\infty$ and will be used even to provide an L^∞-bound on $\phi^\epsilon - c_\epsilon$.

The next step is the following lemma.

LEMMA 2.4. *There exists a continuous, space-time periodic solution of* (1.4).

Proof. 1. Define the function

$$H^K(x,t,p) = \begin{cases} H(x,t,p) & \text{if } |p| \le K, \\ \\ H\Big(x,t,K|p|^{-1}p\Big) & \text{if } |p| \ge K, \end{cases}$$

where K is the constant given by Lemma 2.2. It follows that H^K satisfies (H1) with the same function χ as H.

2. Consider the problem

$$\phi_t^{\epsilon,K} - \Delta\phi^{\epsilon,K} + H^K(x,t,D\phi^{\epsilon,K}) + \epsilon\phi^{\epsilon,K} = 0 \quad \text{in } \mathbb{R}^N \times \mathbb{R}, \tag{2.4}$$

which can be seen as a stationary equation set in $\mathbb{R}^N \times \mathbb{R}$.

If $M = \|H^K(x,t,0)\|_\infty = \|H(x,t,0)\|_\infty$, then $-\epsilon^{-1}M$ and $\epsilon^{-1}M$ are, respectively, sub- and supersolutions of (2.4). Moreover, the special dependence of H^K in p allows for a comparison result between discontinuous sub- and supersolutions of (2.4). Since the supremum of space-time periodic subsolutions is itself space-time periodic, Perron's method (see the user's guide by Crandall, Ishii, and Lions [7]) leads to the existence of a space-time periodic solution of (2.4) satisfying

$$-\epsilon^{-1}M \le \phi^{\epsilon,K}(x,t) \le \epsilon^{-1}M \ \text{ in } \mathbb{R}^N \times \mathbb{R}. \tag{2.5}$$

We conclude by observing that Lemma 2.2 clearly implies that $\phi^\epsilon = \phi^{\epsilon,K}$ is a solution of (2.1) as well.

3. In order to let $\epsilon \to 0$, set $\lambda_\epsilon = \epsilon\phi^\epsilon(0,0)$ and consider the function

$$\underline{\phi}^\epsilon(x,t) = \phi^\epsilon(x,t) - \phi^\epsilon(0,0),$$

which solves the initial value problem

$$\begin{cases} \underline{\phi}_t^\epsilon - \Delta\underline{\phi}^\epsilon + H(x,t,D\underline{\phi}^\epsilon) + \epsilon\underline{\phi}^\epsilon + \lambda_\epsilon = 0 & \text{in } \mathbb{R}^N \times (0,+\infty), \\ \\ \underline{\phi}^\epsilon(x,0) = \phi^\epsilon(x,0) - \phi^\epsilon(0,0) & \text{in } \mathbb{R}^N. \end{cases} \tag{2.6}$$

The gradient bound and the space-time periodicity of ϕ^ϵ implies an independent of ϵ, L^∞-bound on $\underline{\phi}^\epsilon(x,0)$, and this together with (2.5)—which provides the boundedness of λ_ϵ—yields the existence of subsequences, which, for notational simplicity, are also denoted by ϵ such that $\lambda_\epsilon \to \lambda \in \mathbb{R}$ and $\underline{\phi}^\epsilon(x,0)$ converges uniformly in $\mathbb{R}^N$.

Since ϕ^ϵ is Lipschitz continuous in x, the comparison result for (2.6) yields, for any $\epsilon_1, \epsilon_2 > 0$,

$$\sup_{0\le t\le T} \|\underline{\phi}^{\epsilon_1}(\cdot,t) - \underline{\phi}^{\epsilon_2}(\cdot,t)\|_\infty \le \|\underline{\phi}^{\epsilon_1}(x,0) - \underline{\phi}^{\epsilon_2}(x,0)\|_\infty + T|\lambda_{\epsilon_1} - \lambda_{\epsilon_2}|.$$

Hence the family $(\underline{\phi}^\epsilon)_{\epsilon>0}$ converges uniformly in $\mathbb{R}^N \times [0,T]$ to ϕ, which, in view of the stability result for viscosity solutions, is a solution of

$$\phi_t - \Delta\phi + H(x,t,D\phi) + \lambda = 0 \ \text{ in } \mathbb{R}^N \times (0,T).$$

Moreover, ϕ is x-periodic with $\phi(x,T) = \phi(x,0)$ for any $x \in \mathbb{R}^N$. Extending ϕ for $t \in \mathbb{R}$ by periodicity provides the desired solution. ∎

2.2. The superlinear case.

The main assumption here is that

$$\text{(H2)}\quad \begin{cases} H \in C(\mathbb{R}^N \times \mathbb{R} \times \mathbb{R}^N) \text{ is } [0,\Lambda]^N \times [0,T] \text{ space-time periodic in } \mathbb{R}^N \times \mathbb{R}. \\ \text{Moreover, for all } t \in \mathbb{R},\ (x,p) \mapsto H(x,t,p) \text{ is locally Lipschitz continuous} \\ \text{on } \mathbb{R}^N \times \mathbb{R}^N, \text{ and there exists } L > 0 \text{ such that, for all } |p| \geq L, x \in \mathbb{R}^N, \\ t \in \mathbb{R}, \\ \qquad (H_p \cdot p - H - \|H(x,t,0)\|_\infty)L - |H_x| \geq 0. \end{cases}$$

Assumption (H2) is typically satisfied by a nonlinearity H of the form

$$H(x,t,p) = a(x,t)|p|^{1+\beta} + f(x,t),$$

where $\beta > 0$ (superlinear growth), $a, f \in C(\mathbb{R}^N \times \mathbb{R}) \cap L^\infty(\mathbb{R}, W^{1,\infty}(\mathbb{R}^N))$ are $[0,\Lambda]^N \times [0,T]$ space-time periodic in $\mathbb{R}^N \times \mathbb{R}$, and there exists $\eta > 0$ such that

$$a \geq \eta > 0 \quad \text{on } \mathbb{R}^N \times \mathbb{R}.$$

The result is the following theorem.

THEOREM 2.5. *The results of Theorem* 2.1 *hold if we replace* (H1) *by* (H2).

The proof of Theorem 2.5 follows the structure of the proof of Theorem 2.1. The only difference is the way that the gradient bound on the solution of (2.1) is obtained. Indeed in Theorem 2.1 the gradient bound clearly came from the Laplacian, while here it will come from the H term.

The key is the following lemma.

LEMMA 2.6. *If* (H2) *holds, ϕ^ϵ is a continuous, space-time periodic solution of* (2.1)*, and $w^\epsilon : \mathbb{R}^N \times \mathbb{R} \to \mathbb{R}$ is defined by*

$$\exp(w^\epsilon) = \phi^\epsilon - \min_{\mathbb{R}^N \times \mathbb{R}} \phi^\epsilon + 1,$$

then, for the L given in (H2)*,*

$$\sup_{t \in \mathbb{R}} \|Dw^\epsilon\|_{L^\infty(\mathbb{R}^N)} \leq L.$$

We continue with a brief sketch of the proof of Theorem 2.5 using Lemma 2.6.

Proof of Theorem 2.5. We follow the proof of Theorem 2.1 and, in particular, of Lemma 2.4 with the modifications we describe below.

Since ϕ^ϵ is continuous and space-time periodic, there exists (x_ϵ, t_ϵ) such that

$$\phi^\epsilon(x_\epsilon, t_\epsilon) = \min_{\mathbb{R}^N \times \mathbb{R}} \phi^\epsilon.$$

Set $\lambda_\epsilon = \epsilon \phi^\epsilon(x_\epsilon, t_\epsilon)$ and consider the function

$$\underline{\phi}^\epsilon(x,t) = \phi^\epsilon(x,t) - \phi^\epsilon(x_\epsilon, t_\epsilon)\ ,$$

which solves the initial value problem

$$\text{(2.7)}\quad \begin{cases} \underline{\phi}^\epsilon_t - \Delta \underline{\phi}^\epsilon + H(x,t,D\underline{\phi}^\epsilon) + \epsilon \underline{\phi}^\epsilon + \lambda_\epsilon = 0 & \text{in } \mathbb{R}^N \times (t_\epsilon, t_\epsilon + T), \\ \underline{\phi}^\epsilon(x,t_\epsilon) = \phi^\epsilon(x,t_\epsilon) - \phi^\epsilon(x_\epsilon, t_\epsilon) & \text{on } \mathbb{R}^N. \end{cases}$$

Since $w^\epsilon(x_\epsilon, t_\epsilon) = 0$, the gradient bound of Lemma 2.6 and the space periodicity of w^ϵ yield, independent of ϵ, an L^∞-bound for $w^\epsilon(\cdot, t_\epsilon)$, and, therefore, for $\underline{\phi}^\epsilon(\cdot, t_\epsilon)$.

Moreover, the comparison result for (2.7) yields

$$\sup_{t_\epsilon \le t \le t_\epsilon + T} \|\underline{\phi}(\cdot, t)\|_\infty \le \|\underline{\phi}^\epsilon(\cdot, t_\epsilon)\|_\infty + T(M + |\lambda_\epsilon|),$$

where M is defined as in the proof of Lemma 2.4.

Since $|\lambda_\epsilon| \le M$ and because $\underline{\phi}^\epsilon$ is time-periodic, the inequality above provides, independent of ϵ and t, an L^∞-bound for $\underline{\phi}^\epsilon$ and w^ϵ which, together with the bound on Dw^ϵ, also implies such a bound for $\|D\underline{\phi}^\epsilon\|_\infty$.

Finally, the existence of ϕ^ϵ and ϕ follows from the same standard approximation arguments as in the proof of Lemma 2.4. □

Proof of Lemma 2.6. 1. The function w^ϵ solves

$$w_t^\epsilon - \Delta w^\epsilon - |Dw^\epsilon|^2 + b(x, t, w^\epsilon, Dw^\epsilon) = 0 \quad \text{in } \mathbb{R}^N \times \mathbb{R},$$

where

$$b(x, t, u, p) = \exp(-u) H(x, t, \exp(u)p) + \epsilon + \epsilon(m_\epsilon - 1)\exp(-u)$$

and

$$m_\epsilon = \min_{\mathbb{R}^N \times \mathbb{R}} \phi^\epsilon.$$

2. Next consider

$$\max_{\mathbb{R}^N \times \mathbb{R}} (w^\epsilon(x, t) - w^\epsilon(y, t) - L|x - y|),$$

where L is given by (H2). By the same arguments as in the proof of Lemma 2.2, this is indeed a maximum, since w^ϵ is space-time periodic and continuous, it is achieved at $(\overline{x}, \overline{y}, \overline{t})$, and it is nonnegative. If the max is zero, the desired gradient bound follows.

3. If the max is positive, then $\overline{x} \neq \overline{y}$. As in the proof of Lemma 2.2, there exist $a \in \mathbb{R}$ and symmetric matrices X, Y such that $(a, p, X) \in D^{2,+} w^\epsilon(\overline{x}, \overline{t})$ and $(a, p, Y) \in D^{2,-} w^\epsilon(\overline{y}, \overline{t})$ with $p = |\overline{x} - \overline{y}|^{-1} L(\overline{x} - \overline{y})$.

Subtracting the viscosity inequalities

$$a - X - |p|^2 + b(\overline{x}, \overline{t}, w^\epsilon(\overline{x}, \overline{t}), p) \le 0 \quad \text{and} \quad a - Y - |p|^2 + b(\overline{y}, \overline{t}, w^\epsilon(\overline{y}, \overline{t}), p) \ge 0$$

and using that $X \le Y$ yields

$$b(\overline{x}, \overline{t}, w^\epsilon(\overline{x}, \overline{t}), p) - b(\overline{y}, \overline{t}, w^\epsilon(\overline{y}, \overline{t}), p) \le 0.$$

4. To simplify the presentation, we argue next as if H and, therefore, b were C^1 functions. A complete, rigorous proof can be made by standard regularization and approximation arguments.

For $s \in [0, 1]$, set

$$X_s = (s\overline{x} + (1 - s)\overline{y}, \overline{t}, s w^\epsilon(\overline{x}, \overline{t}) + (1 - s) w^\epsilon(\overline{y}, \overline{t}), p)$$

and rewrite the last inequality as

$$b(X_1) - b(X_0) = \int_0^1 \frac{d}{ds}[b(X_s)] ds \le 0 \,,$$

i.e.,

$$\int_0^1 [b_x(X_s)\cdot(\overline{x}-\overline{y}) + b_u(X_s)(w^\epsilon(\overline{x},\overline{t}) - w^\epsilon(\overline{y},\overline{t}))]ds \le 0\ . \tag{2.8}$$

Straightforward computations yield

$$b_x(x,t,u,p) = \exp(-u)H_x(x,t,\exp(u)p)$$

and

$$b_u(x,t,u,p) = \exp(-u)\left[H_p\cdot p - H\right](x,t,\exp(u)p) - \epsilon(m_\epsilon - 1)\exp(-u)\ .$$

5. Since $\|\epsilon\phi^\epsilon\|_{L^\infty(\mathbb{R}^N\times\mathbb{R})} \le \|H(x,t,0)\|_\infty$, it follows that $\epsilon m_\epsilon \le \|H(x,t,0)\|_\infty$ and, therefore,

$$b_u(X_s) > \exp(-u)\Big[H_p\cdot p - H - \|H(x,t,0)\|_\infty\Big](x,t,\exp(u)p).$$

6. Since by definition $w^\epsilon \ge 0$, we have

$$u = sw^\epsilon(\overline{x},\overline{t}) + (1-s)w^\epsilon(\overline{y},\overline{t}) \ge 0,$$

hence

$$\exp(u)|p| \ge L, \tag{2.9}$$

and, in view of (H2),

$$b_u(X_s) > 0.$$

Finally, the assumption that the max is positive gives

$$w^\epsilon(\overline{x},\overline{t}) - w^\epsilon(\overline{y},\overline{t}) > L|\overline{x}-\overline{y}|.$$

Combining all the above information we obtain

$$b_u(X_s)(w^\epsilon(\overline{x},\overline{t})-w^\epsilon(\overline{y},\overline{t})) > \exp(-u)\big[H_p\cdot p - H - \|H(x,t,0)\|_\infty\big](x,t,\exp(u)p)L|\overline{x}-\overline{y}|.$$

7. If

$$Q = [b_x(X_s)\cdot(\overline{x}-\overline{y}) + b_u(X_s)(w^\epsilon(\overline{x},\overline{t}) - w^\epsilon(\overline{y},\overline{t}))]\ ,$$

then

$$\begin{aligned} Q &> \exp(-u)|\overline{x}-\overline{y}|\left[H_x\cdot\frac{\overline{x}-\overline{y}}{|\overline{x}-\overline{y}|} + (H_p\cdot p - H - \|H(x,t,0)\|_\infty)L\right] \\ &> \exp(-u)|\overline{x}-\overline{y}|\left[(H_p\cdot p - H - \|H(x,t,0)\|_\infty)L - |H_x|\right]. \end{aligned}$$

Using (H2) and recalling (2.9) yields the strict positivity of the integral of (2.8), which provides the desired contradiction. □

Remark 2.7. The argument used in the proof of Lemma 2.6 follows along the lines of Barles [1], which introduced a "weak" Bernstein method to obtain gradient bounds on solutions of fully nonlinear PDEs. Here the difference is that L^∞-bounds on the w^ϵ cannot be used, and this is a nontrivial additional difficulty.

3. Extensions to quasi-linear equations. In this section we provide two existence results of continuous, space-time periodic solutions of (1.1). They correspond to the sub- and superlinear cases presented in the previous section. We omit their proofs, since they are only routine adaptations of the proofs of Theorems 2.1 and 2.5, although they involve more tedious computations. In what follows $\mathcal{M}_{N,M}$ denotes the space of $N \times M$ matrices and $\mathcal{M}_N$ denotes the space of $N \times N$ matrices.

Throughout the section we assume

$$\text{(H3)} \qquad \begin{cases} A \in C(\mathbb{R}^N \times \mathbb{R} \times \mathbb{R}^N, S^N) \text{ and } H \in C(\mathbb{R}^N \times \mathbb{R} \times \mathbb{R}^N) \\ \\ \text{are } [0,\Lambda]^N \times [0,T] \text{ space-time periodic.} \end{cases}$$

For the *sublinear case*, we introduce the following additional hypotheses:

$$\text{(H4)} \qquad \begin{cases} \text{There exists a } \mu : \mathbb{R} \to (0,\infty) \text{ and a constant } C > 0 \text{ such that, for all} \\ (x,t,p) \in \mathbb{R}^N \times \mathbb{R} \times \mathbb{R}^N, \\ \\ \qquad C^{-1}\mu(|p|)\mathrm{Id} \le A(x,t,p) \le C\mu(|p|)\mathrm{Id}, \end{cases}$$

$$\text{(H5)} \qquad \begin{cases} \text{there exist } \sigma \in C(\mathbb{R}^N \times \mathbb{R} \times \mathbb{R}^N; \mathcal{M}_N),\ \theta \in \left(\tfrac{1}{2}, 1\right], \text{ and } C > 0 \text{ such that} \\ \\ \qquad A = \sigma\sigma^T \text{ and } |\sigma(x,t,p) - \sigma(y,t,p)| \le C\mu^{1/2}(|p|)|x-y|^\theta, \end{cases}$$

and

$$\text{(H6)} \qquad \begin{cases} \text{there exists } \chi \in C(\mathbb{R}^+, \mathbb{R}^+) \text{ such that } \displaystyle\int^{+\infty} \chi(u)^{-1} du = +\infty, \text{ and} \\ \\ |H(x,t,p)| \le \mu(|p|)\chi(|p|) \text{ for all } (x,t,p) \in \mathbb{R}^N \times \mathbb{R} \times \mathbb{R}^N. \end{cases}$$

Our result is the following theorem.

THEOREM 3.1. *Assume that* (H3)–(H6) *hold. Then there exists a solution* (ϕ, λ) *of* (1.1)*, where* ϕ *is a continuous space-time periodic function. Moreover, the constant* λ *is unique.*

The proof of Theorem 3.1 follows exactly along the lines of the one of Theorem 2.1. Assumption (H5) is used both to get a gradient bound on the ϕ^ϵ and to have a comparison result available for the PDE satisfied by ϕ^ϵ. The gradient bound by itself can be obtained with weaker assumptions. We refer to Barles [2] for discussions in this direction.

For the *superlinear case* we introduce the following assumptions:

$$\text{(H7)} \qquad \begin{cases} \text{There exists a locally Lipschitz continuous function} \\ \\ \sigma : \mathbb{R}^N \times \mathbb{R} \times \mathbb{R}^N \to \mathcal{M}_{N,p} \text{ such that } A = \sigma\sigma^T, \end{cases}$$

and

$$\text{(H8)} \qquad \begin{cases} H \text{ is locally Lipschitz continuous with respect to } (x,p) \text{ for all } t \in \mathbb{R}, \\ H_p \cdot p - H \to +\infty \text{ as } |p| \to +\infty \text{ uniformly with respect to } (x,t), \\ \text{and, as } |p| \to \infty, \\ \\ \qquad A_p \cdot p|p|^2,\ (\sigma_p \cdot p)^2|p|^2 = o(H_p \cdot p - H), \\ \\ \qquad (A_x \cdot p)^2,\ (A_x \cdot p)|p|,\ H_x \cdot p = o((H_p \cdot p - H)|p|). \end{cases}$$

THEOREM 3.2. *Under assumptions* (H3), (H7), *and* (H8), *the conclusions of Theorem* 2.1 *remain valid.*

In Theorem 3.2, the gradient estimate is obtained following the method of proof (or by using directly the result) of Barles [1, Theorem 2, p. 265]) after an exponential change of variable is performed, as in the proof of Theorem 2.5. It is necessary, however, to control carefully the dependence of all the terms in w^ϵ since, again, we have no a priori L^∞-bound on w^ϵ.

4. Convergence as $t \to \infty$ to space-time periodic solutions.

Our main result is the following theorem.

THEOREM 4.1. *Under the assumptions of Theorem* 3.1 *or* 3.2, *for any* $u_0 \in W^{1,\infty}(\mathbb{R}^N)$ $[0,\Lambda]^N$*-periodic in* x, *there exists a unique solution* u *of* (1.2), *which is in* $W^{1,\infty}(\mathbb{R}^N)$ *for any* $t > 0$ *and* $[0,\Lambda]^N$*-periodic in* x. *Moreover,* $u(x,t) - \lambda t$ *is uniformly bounded. If* (1.2) *satisfies the strong maximum principle, then* (1.3) *holds for some* $[0,\Lambda]^N \times [0,T]$*-space-time periodic solution of* (1.1).

We will return to the assumption that (1.2) satisfies the strong maximum principle after the short proof of Theorem 4.1. We want, however, just to point out here that this assumption implies the uniqueness of ϕ up to constants.

Proof. 1. The existence of u relies essentially on the same arguments as the existence of the space-time periodic solution ϕ of (1.1). The uniqueness uses in an essential way the boundedness of Du.

2. Since $u(x,t) - \lambda t$ and ϕ are solutions in $\mathbb{R}^N \times (0,+\infty)$ of the same equation, a standard comparison principle yields

$$\|u(x,t) - \lambda t - \phi(x,t)\|_\infty \le \|u_0(x) - \phi(x,0)\|_\infty,$$

and, therefore, $u(x,t) - \lambda t$ remains bounded in L^∞. This also leads to an easier gradient bound for $u(x,t) - \lambda t$.

3. Setting $v(x,t) = u(x,t) - \lambda t$, applying the same comparison result on $\mathbb{R}^N \times (t,+\infty)$ instead of $\mathbb{R}^N \times (0,\infty)$, and using that ϕ is T-periodic in time yields, for $s \ge t$,

$$\max_x (v(x,s) - \phi(x,s)) \le \max_x (v(x,t) - \phi(x,t)).$$

Therefore,

$$m(t) = \max_x (v(x,t) - \phi(x,t))$$

is decreasing, and, since $m(t)$ is bounded,

$$m(t) \to \overline{m} \quad \text{as } t \to +\infty.$$

4. The sequence $(v(\cdot,kT))_k$ is compact in $C(\mathbb{R}^N)$. Therefore, using the periodicity in x, it is possible to extract a subsequence, which for notational simplicity is also denoted $(v(\cdot,kT))_k$, converging uniformly to $\overline{v} \in W^{1,\infty}(\mathbb{R}^N)$. Moreover, the comparison result for viscosity solution yields

$$\|v(\cdot,t+kT) - v(\cdot,t+\widehat{k}T)\|_\infty \le \|v(\cdot,kT) - v(\cdot,\widehat{k}T)\|_\infty \to 0 \quad \text{as } k,\widehat{k} \to +\infty.$$

Hence $v(x,t+kT) \to w(x,t)$, which is, by the stability of viscosity solution, a solution of

$$(4.1) \quad \begin{cases} w_t - \operatorname{tr}(A(x,t,Dw)D^2w) + H(x,t,Dw) + \lambda = 0 & \text{in } \mathbb{R}^N \times (0,+\infty), \\ w(x,0) = \overline{v}(x) & \text{in } \mathbb{R}^N. \end{cases}$$

5. Passing to the limit in

$$m(t+kT) = \max_x (v(x,t+kT) - \phi(x,t))$$

and using the uniform convergence of the $v(\cdot,\cdot+kT)$ yields

$$\overline{m} = \max_x (w(x,t) - \phi(x,t)) \quad \text{for any } t > 0.$$

The strong maximum principle then implies that

$$w = \phi + \overline{m}.$$

An easy argument then gives (1.3) with ϕ replaced by $\phi + \overline{m}$. Indeed, on one hand, since $\overline{v} = \phi + \overline{m}$ is the only possible limit of any subsequence of the sequence $(v(\cdot,kT))_k$, it follows that $v(\cdot,kT) \to \phi(\cdot,0) + \overline{m}$. On the other hand, for t large, we consider the integer k such that $kT \le t < (k+1)T$. A standard comparison result together with the time periodicity of ϕ yields

$$\begin{aligned} \|u(x,t) - \lambda t - [\phi(x,t) + \overline{m}]\|_\infty &= \|v(x,t) - [\phi(x,t) + \overline{m}]\|_\infty \\ &\le \|v(x,kT) - \phi(x,kT)\|_\infty \\ &= \|v(x,kT) - [\phi(x,0) + \overline{m}]\|_\infty \to 0 \\ &\quad \text{as } t \to +\infty. \qquad \square \end{aligned}$$

Next we turn to the question, When is the strong maximum principle satisfied by (4.1)? In fact there are two very different cases, depending on whether A depends on p or not.

When A *does not depend on* p, the strong maximum principle can easily be established, using rather soft arguments, in the general framework of viscosity solutions. In fact, in addition to the uniform ellipticity of the equation, it requires only the basic assumptions of a classical comparison result together with a suitable hypothesis on the Lipschitz continuity of H with respect to p. More precise assumptions will be given below. It is, however, worth pointing out that, since we deal here with solutions which are Lipschitz continuous in x, these assumptions are rather weak. In particular, in this case, no deep regularity result on the solutions is needed and as a consequence the strong maximum principle can be applied to equations where the solutions are not classical.

On the contrary, if A *depends on* p, to the best of our knowledge, the strong maximum principle holds only for classical solutions, and therefore a regularity result for the solutions of (4.1) is needed. The form of the equation and the classical parabolic theory suggest that it is enough to have solutions which are $C^{1,\alpha}$ in x. Indeed then Schauder's theory yields $C^{2,1}$ solutions for (4.1), where $C^{2,1}$ is the space of functions which are C^2 in x and C^1 in t. The strong maximum principle then follows.

There has been a substantial body of work about the question of $C^{1,\alpha}$ regularity of solutions of fully nonlinear, uniformly elliptic/parabolic PDEs (see, for example, Caffarelli [4], Caffarelli and Cabré [5], Crandall et al. [6], Krylov [11], Wang [21], [22], etc.). There are basically two approaches. The first uses difference quotients for viscosity solutions and yields $C^{1,\alpha}$ bounds. The required assumptions do not, however, cover quasi-linear equations, when A depends on Dw. The second approach deals with general fully nonlinear problems and yields $C^{1,\alpha}$ bounds for smooth, say, C^3-solutions, for which there is not, however, in general, existence.

It turns out that for (4.1) under the assumptions (H3)–(H6) or (H3), (H7), and (H8) which yield a priori Lipschitz estimates, the gap described above can be closed when (H4) holds. Indeed, as soon as one has a priori bounds on the gradient, the theory developed in Chapter VI of Ladyzenskaja, Solonnikov, and Ural'ceva [12] provides a $C^{1,\alpha}$-bound on the solutions which then yields a $C^{2,\alpha}$-bound by the classical regularity theory. The existence of a space-time periodic, classical solution then follows from a straightforward truncation of A and H. Moreover, this solution is uniformly bounded and even compact in $C^{2,1}$ equipped with the natural norm.

Now we come back to the original convergence problem. In the statement of our result below, in the case when A is independent of p, we will say that A satisfies assumptions (H′4) and (H′5) if (H4) and (H5) hold with σ being independent of p and $\mu(|p|)$ being replaced by a constant $\tilde{\mu} > 0$.

The above discussion on the strong maximum principle and the existence of $C^{2,1}$-solutions together with Theorem 4.1 leads to the following theorem.

THEOREM 4.2. *Assume that either* (i) *A is independent of p and satisfies* (H3), (H′4), (H′5) *and that H is locally Lipschitz continuous with respect to p, uniformly in x and t, or* (ii) *A and H are locally Lipschitz continuous functions satisfying* (H3) *and* (H4). *If the assumptions of Theorem* 3.1 *or* 3.2 *are satisfied, then, for any $u_0 \in W^{1,\infty}(\mathbb{R}^N)$ which is $[0,\Lambda]^N$-periodic in x, there exists a unique solution u of* (1.2), *which is in $W^{1,\infty}(\mathbb{R}^N)$ for any $t > 0$, is $[0,\Lambda]^N$-periodic in x, and* (1.3) *holds for some $[0,\Lambda]^N \times [0,T]$-space-time periodic solution of* (1.1).

REFERENCES

[1] G. BARLES, *A weak Bernstein method for fully nonlinear elliptic equations*, Differential Integral Equations, 4 (1991), pp. 241–262.

[2] G. BARLES, *Interior gradient bounds for the mean curvature equation by viscosity solutions methods*, Differential Integral Equations, 4 (1991), pp. 263–275.

[3] G. BARLES AND P. E. SOUGANIDIS, *Counterexamples on the Asymptotic Behavior of the Solutions of Hamilton–Jacobi Equations*, preprint.

[4] L. A. CAFFARELLI, *Interior a priori estimates for solutions of fully nonlinear equations*, Ann. of Math. (2), 130 (1989), pp. 189–213.

[5] L. A. CAFFARELLI AND X. CABRÉ, *Fully Nonlinear Elliptic Equations*, Amer. Math. Soc. Colloq. Publ. 43, AMS, Providence, RI, 1995.

[6] M. G. CRANDALL, K. FOK, M. KOCAN, AND A. ŚWIECH, *Remarks on nonlinear uniformly parabolic equations*, Indiana Univ. Math. J., 47 (1998), pp. 1293–1326.

[7] M. G. CRANDALL, H. ISHII, AND P.-L. LIONS, *User's guide to viscosity solutions of second order partial differential equations,* Bull. Amer. Math. Soc. (N.S.), 27 (1992), pp. 1–67.

[8] A. FATHI AND J. N. MATHER, *Failure of Convergence of the Lax–Oleinik Semi-group in the Time-Periodic Case*, preprint.

[9] M. I. FREIDLIN, *Functional Integration and Partial Differential Equations,* Ann. of Math. Stud. 109, Princeton University Press, Princeton, NJ, 1985.

[10] H. ISHII AND P.-L. LIONS, *Viscosity solutions of fully nonlinear second-order elliptic partial differential equations*, J. Differential Equations, 83 (1990), pp. 26–78.

[11] N. KRYLOV, *Nonlinear Elliptic and Parabolic Equations of the Second Order*, Reidel Publishing, Dordrecht, The Netherlands, 1987.

[12] O. A. LADYZENSKAJA, V. A. SOLONNIKOV, AND N. N. URAL'CEVA, *Linear and Quasilinear Equations of Parabolic Type*, Transl. Math. Monogr. 23, AMS, Providence, RI, 1967.

[13] P. L. LIONS, *Neumann type boundary conditions for Hamilton-Jacobi Equations*, Duke Math. J., 52 (1985), pp 793–820.

[14] P.-L. LIONS, G. PAPANICOLAOU, AND S. R. S. VARADHAN, *Homogenization of Hamilton-Jacobi Equations*, manuscript.

[15] A. MAJDA AND P. E. SOUGANIDIS, *Large scale front dynamics for turbulent reaction-diffusion equations with separated velocity scales*, Nonlinearity, 7 (1994), pp. 1–30.

[16] G. NAMAH AND J. M. ROQUEJOFFRE, *Remarks on the long time behavior of the solutions of Hamilton–Jacobi Equations*, Comm. Partial Differential Equations, to appear.

[17] G. Namah and J. M. Roquejoffre, *Convergence to Periodic Fronts in a Class of Semilinear Parabolic Equations*, preprint.
[18] J. M. Roquejoffre, *Comportement asymptotique des solutions d'équations de Hamilton–Jacobi monodimensionnelles*, C. R. Acad. Sci. Paris Sér. I Math., 326 (1998), pp. 185–189.
[19] J. M. Roquejoffre, *Convergence to Steady States or Periodic Solutions in a Class of Hamilton–Jacobi Equations*, preprint.
[20] Ya. G. Sinai, preprint.
[21] L. Wang, *On the regularity of fully nonlinear parabolic equations, Part* I, Comm. Pure Appl. Math., 45 (1992), pp. 27–76.
[22] L. Wang, *On the regularity of fully nonlinear parabolic equations, Part* II, Comm. Pure Appl. Math., 45 (1992), pp. 141–178.

SIAM J. MATH. ANAL.
Vol. 32, No. 6, pp. 1324–1326

A REMARK ON ISOPOLAR POTENTIALS*

MACIEJ ZWORSKI†

Abstract. We discuss the inverse problem for resonances of one-dimensional potentials.

Key words. resonances, inverse problems

AMS subject classifications. 34L25, 81U40

PII. S0036141000375585

1. In the approach to scattering by compactly supported perturbations introduced by Lax and Phillips [3], the scattering poles, or resonances, might be considered as the natural analogue of the discrete spectrum of operators on compact manifolds. One could then ask to what extent could the resonances determine the scatterer and consider the "isopolar" rather than the isospectral problem. I would like to make a simple remark concerning the Schrödinger operator on the line.

Different situations could also be considered. In hyperbolic scattering, using a modification of Sunada's construction, Zelditch [7] and Bérard [1] obtained examples of "isopolar manifolds." We recall that in the finite volume case, hyperbolic scattering can be considered as a type of scattering by a compactly supported perturbation of a half-line—see, for instance, Example 1.1 in [5] and the references given therein.

2. We consider the Schrödinger operator

$$-\frac{d^2}{dx^2}+p(x), \quad x\in\mathbb{R}.$$

We shall follow the notation from [4]. Our conclusion is a simple consequence of standard facts about inverse scattering. We shall be concerned with real even potential p in $L^1_{\rm comp}(\mathbb{R})$ which is equivalent to considering scattering on a half-line, or the lowest angular momentum scattering by a radially symmetric potential in $\mathbb{R}^3$. The scattering matrix has the form

$$S_p(\lambda)=\begin{bmatrix} i\lambda/\widehat{X}(\lambda) & \widehat{Y}(\lambda)/\widehat{X}(\lambda)\\ \widehat{Y}(-\lambda)/\widehat{X}(\lambda) & i\lambda/\widehat{X}(\lambda)\end{bmatrix}.$$

The potential p is said to have a half-bound state at 0 if $\widehat{X}(0)=0$. The scattering poles, or resonances, are given by the zeros of $\widehat{X}$ in $\Im\lambda<0$, but we will also include in that set the square roots of the negative eigenvalues of $-d^2/dx^2+p(x)$, which correspond to the zeros of $\widehat{X}$ in $\Im\lambda>0$. Other than at $\lambda=0$, $\widehat{X}$ cannot vanish on the real axis.

PROPOSITION. *Let $p\in L^1_{\rm comp}(\mathbb{R})$ be even. Then, if p does not have a half-bound state at 0, the scattering poles determine p uniquely. If p has a half-bound state, then*

*Received by the editors July 21, 2000; accepted for publication October 18, 2000; published electronically March 28, 2001.

http://www.siam.org/journals/sima/32-6/37558.html

†Department of Mathematics, The Johns Hopkins University, Baltimore, MD 21218. Current address: Mathematics Department, University of California, Berkeley, CA 94720 (zworski@math.berkeley.edu).

there exists precisely one $p_1 \in L^1_{\text{comp}}(\mathbb{R})$, $p_1 \neq p$, *with the same set of scattering poles.*

Proof. Since the scattering matrix determines a compactly supported potential uniquely (see, for instance, Proposition 8 of [8]) we need to show only that the poles determine $S_p(\lambda)$, i.e., $\widehat{X}$ and $\widehat{Y}$. Let us first assume that $\widehat{X}(0) \neq 0$. We use the facts about the properties of zeros of Fourier transforms from [6], extended to Fourier transforms of distributions as in Lemma 1 of [8]. If the support of $X \in \mathcal{E}'$ is $[-4a, 0]$ (where in fact $2a$ is the diameter of the support of p), then the Carleman formula and the bound $\mathcal{O}(r)$ on the number of poles give, as in Theorem 6 of [6],

$$\widehat{X}(\lambda) = e^{2a\lambda}\widehat{X}(0) \prod_{\text{poles}} \left(1 - \frac{\lambda}{\lambda_j}\right), \quad |\lambda_1| \leq \cdots |\lambda_j| \leq \cdots.$$

The density of the poles is given by $4a/\pi$ so the exponential factor is fully determined. Since $X - \delta' - C\delta \in L^1$, the factor $\widehat{X}(0)$ is determined uniquely—multiplication of $\widehat{X}(0)$ by a constant would destroy the property $\widehat{X}(\lambda) - i\lambda - C \to 0$, $\lambda \to \pm\infty$. It remains to construct $\widehat{Y}(\lambda)$. For that we recall the unitarity relation

$$\widehat{X}(\lambda)\widehat{X}(-\lambda) = \lambda^2 + \widehat{Y}(\lambda)^2, \tag{1}$$

where $\widehat{Y}(\lambda) = \widehat{Y}(-\lambda)$, as p is even. Thus $\widehat{Y}(\lambda) = \pm(-\lambda^2 + \widehat{X}(\lambda)\widehat{X}(-\lambda))^{\frac{1}{2}}$. To determine the branch we recall that the definition of $\widehat{X}$ and $\widehat{Y}$ (see [4] or [8]) implies $\widehat{Y}(0) = -\widehat{X}(0) \neq 0$.

When $\widehat{X}(0) = 0$, the unitarity relation (1) shows that the zero is of order one. Thus we can apply the above argument to $\widehat{X}(\lambda)/\lambda$. Now, however, there is no obstruction to the choice of either branch in the formula for $\widehat{Y}$, and we obtain two different scattering matrices. □

We conclude with a few remarks. The restriction to compactly supported potentials is natural if one wants to define resonances globally. However, if the potential decays faster than any exponential, then $\widehat{X}$ is still holomorphic in $\mathbb{C}$ and again the resonances are globally defined as its zeros. In fact, if $p \in C^\infty(\mathbb{R})$ and $p(x) = \mathcal{O}(\exp(-A|x|^{1+a}))$, then the number of resonances of p in a disc of radius r, $N_p(r)$, satisfies

$$N_p(r) = \mathcal{O}(r^{\frac{1+a}{a}}).$$

This can be seen from the methods of [4] by showing that $X(y) = \delta'(y) - \frac{1}{2}(\int p(z)dz)\delta(y) + \mathcal{O}(\exp(-2^{-2(1+a)}A|y|^{1+a}))$. But, unlike in the compactly supported case (see [8]), the existence of asymptotics and the structure of the Hadamard factorization of $\widehat{X}$ are not clear and one misses up to $[(1+a)/a] + 1$ constants needed for the recovery procedure in the proof of the proposition.

The various methods of constructing potentials with the same scattering matrix (see, for instance, [2]) certainly yield isopolar potentials if one starts with, say, a compactly supported potential (so that the scattering matrix is globally meromorphic), but they immediately destroy the compact support property or, in fact, the super-exponential decay (see Proposition 8 of [8]).

Remark 1. I wrote this paper in 1988 and revised it in 1992. I decided to publish it now in view of renewed interest in one-dimensional resonance problems [9], [11], [12], [13].

For other relevant developments in the theory of resonances I refer the reader to the bibliography of [14] and to [10].

Remark 2. Determining potentials on a half-line from their resonances was recently studied by Kargaev and Korotyaev [12]. Some results on that can also be obtained by the method above: $\hat{X}(\xi)\hat{X}(-\xi) = (-i\xi + \hat{Y}(\xi))(i\xi + \hat{Y}(\xi))$ and the poles for the Neumann and Dirichlet problems are given by zeros of the two factors on the right, respectively. Proceeding as in the proof of the proposition shows that we can determine $\hat{Y}$ from either set, and hence the zeros of $\hat{X}$, and hence $\hat{X}$ itself. The additional information of distinguishing Dirichlet and Neumann spectra now eliminates the ambiguity in the case of a zero resonance. Amusingly, when a zero resonance is present, then the two potentials in the proposition above have their Dirichlet and Neumann resonances interchanged.

Kargaev and Korotyaev also pointed out a mistake in [8, Proposition 8]: the compactness of the support of a potential cannot be concluded from the transmission coefficient alone. That can, however, be done when the potential is even, and that is all that is relevant in this note.

REFERENCES

[1] P. Bérard, *Transplantation et isospecralité* I, Math. Ann., 292 (1992), pp. 547–559.
[2] P.A. Deift, *Application of a commutation formula*, Duke Math. J., 45 (1978), pp. 267–310.
[3] P. Lax and R. Phillips, *Scattering Theory*, Academic Press, New York, 1967.
[4] A. Melin, *Operator methods for inverse scattering on the real line*, Comm. Partial Differential Equations, 10 (1985), pp. 677–766.
[5] J. Sjöstrand and M. Zworski, *Distribution of scattering poles near the real axis*, Comm. Partial Differential Equations, 17 (1992), pp. 1021–1035.
[6] E.C. Titchmarsh, *The zeros of certain integral functions*, Proc. London Math. Soc., 25 (1926), pp. 283–302.
[7] S. Zelditch, *Kuznecov formulæ and Szego limit formulæ on manifolds*, Comm. Partial Differential Equations, 17 (1992), pp. 221–260.
[8] M. Zworski, *Distribution of poles for scattering on the real line*, J. Funct. Anal., 73 (1987), pp. 277–296.
[9] R. Froese, *Asymptotic distribution of resonances in one dimension*, J. Differential Equations, 137 (1997), pp. 251–272.
[10] A. Hassell and M. Zworski, *Resonant rigidity of* S^2, J. Funct. Anal., 169 (1999), pp. 604–609.
[11] M. Hitrik, *Bounds on scattering poles in one dimension*, Comm. Math. Phys., 208 (1999), pp. 381–411.
[12] P. Kargaev and E. Korotyaev, *Inverse Resonance Scattering and Conformal Mappings for Schrödinger Operators*, preprint, 2000.
[13] B. Simon, *Resonances in One Dimension and Fredholm Determinants*, preprint, 2000.
[14] M. Zworski, *Resonances in physics and geometry,* Notices Amer. Math. Soc., 46 (1999), pp. 319–328.

SIAM J. MATH. ANAL.
Vol. 32, No. 6, pp. 1327–1341

EXPLICIT CHARACTERIZATION OF INCLUSIONS IN ELECTRICAL IMPEDANCE TOMOGRAPHY*

MARTIN BRÜHL†

Abstract. In electrical impedance tomography one seeks to recover the spatial conductivity distribution inside a body from knowledge of the Neumann–Dirichlet map. In many practically relevant situations the conductivity is smooth apart from some inhomogeneities where the conductivity jumps to a higher or lower value. An explicit characterization of these inclusions is developed in this paper. To this end a class of dipole-like indicator functions is introduced, for which one has to check whether their boundary values are contained in the range of an operator determined by the measured Neumann–Dirichlet map. It is shown that this holds true if and only if the dipole singularity lies inside the inhomogeneity. This procedure is conceptually similar to a recent method proposed by Kirsch in inverse scattering theory.

Key words. electrical impedance tomography, inverse boundary value problem

AMS subject classifications. 35R30, 31A25, 31B20, 35R05

PII. S003614100036656X

1. Introduction.

In 1980 Calderón [4] posed the following inverse boundary value problem: is the conductivity coefficient $\sigma(x) > 0$ in the elliptic equation

$$\nabla \cdot \sigma \nabla u = 0 \qquad \text{in } B \tag{1.1}$$

uniquely determined from knowledge of the Neumann and Dirichlet boundary values of all its solutions u, and, if yes, how can σ be reconstructed from these data? Here, B denotes a two- or three-dimensional domain with smooth boundary $T = \partial B$. The first of these questions has been answered in the affirmative for broad classes of conductivities. Kohn and Vogelius [14], Sylvester and Uhlmann [20], Nachman [17], and Brown and Uhlmann [1] proved uniqueness under certain smoothness assumptions on σ. For discontinuous but piecewise smooth conductivities uniqueness has been settled by Kohn and Vogelius [15] and Isakov [12].

This work is devoted to the second question and gives an—at least partial—answer to the reconstruction problem in the case of discontinuous conductivities. For ease of exposition we restrict ourselves to the case of a constant background conductivity, $\sigma = \mathbb{1}$, where $\mathbb{1}$ is the function identically 1. We consider conductivities of the following form:

$$\sigma(x) = \begin{cases} \kappa(x), & x \in \overline{\Omega}, \\ 1, & x \in B \setminus \overline{\Omega}, \end{cases} \tag{1.2}$$

where Ω, $\overline{\Omega} \subset B$, is a collection of separated and simply connected domains with sufficiently smooth boundary $\Gamma = \partial\Omega$, and the conductivity within Ω is significantly higher or lower than the background conductivity, or more precisely

$$\kappa(x) \geq 1 + \varepsilon \qquad \text{or} \qquad \kappa(x) \leq 1 - \varepsilon \qquad \text{for } x \in \overline{\Omega} \text{ and some } \varepsilon > 0. \tag{1.3}$$

*Received by the editors March 10, 2000; accepted for publication (in revised form) October 5, 2000; published electronically March 28, 2001. This work was supported by the Deutsche Forschungsgemeinschaft (DFG).

http://www.siam.org/journals/sima/32-6/36656.html

†Fachbereich Mathematik, Johannes Gutenberg–Universität Mainz, 55099 Mainz, Germany (bruehl@math.uni-mainz.de).

We will refer to the components of Ω as *inclusions*.

In this paper we give an explicit representation of the inclusions Ω in terms of the measured Neumann–Dirichlet operator Λ_σ which maps the Neumann boundary values of potentials u in (1.1) to its Dirichlet boundary values. For that we consider a Neumann function $N(z,x)$ for the Laplace equation in B and, for some direction d, its directional derivative $d\cdot\nabla_z N(z,x)$ as a function of x. Then we show that its boundary values $g_{z,d} = (d\cdot\nabla_z N(z,\cdot))|_T$ are contained in the range $\mathcal{R}(|\Lambda_\sigma - \Lambda_{\mathbb{1}}|^{1/2})$ if and only if the parameter point z lies in Ω. Here, the operators Λ_σ and $\Lambda_{\mathbb{1}}$ are regarded as operators in L^2 topology, and from their compactness it follows that $\mathcal{R}(|\Lambda_\sigma - \Lambda_{\mathbb{1}}|^{1/2})$ is a nonclosed subspace of $L^2(T)$; see section 2 for details.

It should be emphasized that the idea of characterizing the support of an obstacle by the range of some operator has been introduced by Colton and Kirsch [5] and rigorously justified by Kirsch [13] in the context of inverse scattering problems. Recently, Hähner [9] has carried over Kirsch's techniques to the impedance tomography problem for the special case of grounded perfect electrical conductors, i.e., $\kappa = \infty$ in $\overline{\Omega}$. For the general situation addressed in this paper Ikehata [11] has established a characterization of inclusions, which happens to be nonconstructive; on the other hand, in [10] he has indicated a constructive procedure to recover the convex hull of the inclusions.

This paper is organized as follows. In the next section we introduce the notions associated with the forward problem and summarize some properties of the Neumann–Dirichlet operator. In section 3 we state and prove our main result for the case of a single inclusion and piecewise constant conductivity. Generalizations of this special case are treated in section 4.

The subtle difficulties that need to be overcome to turn our constructive method of proof into a numerical algorithm are discussed in [3].

2. Forward problem. For every $f \in H^{-1/2}(T)$ with $\langle f, \mathbb{1}\rangle_{L^2(T)} = \oint_T f\,ds = 0$ the Neumann boundary value problem

$$\nabla\cdot\sigma\nabla u = 0 \quad \text{in } B, \qquad \sigma\frac{\partial u}{\partial \nu} = f \quad \text{on } T \tag{2.1}$$

has a unique weak solution $u_\sigma \in H^1(B)$ with $\oint_T u_\sigma\,ds = 0$. The expression $\langle\phi,\psi\rangle_{L^2(T)} = \oint_T \phi\psi\,ds$ will be used for both the L^2 inner product on T and the dual pairing in $\langle H^{-1/2}(T), H^{1/2}(T)\rangle$. Moreover, let us introduce the following subspaces of the usual Sobolev spaces:

$$\begin{aligned}
L^2_\diamond(T) &= \{\phi \in L^2(T) : \oint_T \phi\,ds = 0\},\\
H^{\pm1/2}_\diamond(T) &= \{\phi \in H^{\pm1/2}(T) : \oint_T \phi\,ds = 0\},\\
L^2_\diamond(\Gamma) &= \{\phi \in L^2(\Gamma) : \oint_\Gamma \phi\,ds = 0\},\\
H^{\pm1/2}_\diamond(\Gamma) &= \{\phi \in H^{\pm1/2}(\Gamma) : \oint_\Gamma \phi\,ds = 0\},\\
H^1_\diamond(B) &= \{u \in H^1(B) : \oint_T u\,ds = 0\},\\
H^1_{\diamond,T}(B\setminus\overline{\Omega}) &= \{u \in H^1(B\setminus\overline{\Omega}) : \oint_T u\,ds = 0\},\\
H^1_{\diamond,\Gamma}(B\setminus\overline{\Omega}) &= \{u \in H^1(B\setminus\overline{\Omega}) : \oint_\Gamma u\,ds = 0\}.
\end{aligned}$$

Then, the solution operator $f \mapsto u_\sigma$ of (2.1) is a well-defined bounded linear operator $H^{-1/2}_\diamond(T) \to H^1_\diamond(B)$ and by passing to the trace of u_σ we obtain the *Neumann–*

Dirichlet operator

$$\Lambda_\sigma : H_\diamond^{-1/2}(T) \longrightarrow H_\diamond^{1/2}(T), \qquad f \longmapsto u_\sigma|_T,$$

which in fact is an isomorphism between these spaces.

The following properties of Λ_σ are well known, see, e.g., [8], but we sketch the proof for convenience.

LEMMA 2.1.

(a) $\Lambda_\sigma : L_\diamond^2(T) \to L_\diamond^2(T)$ *is compact, self-adjoint, and positive.*

(b) Λ_σ *is strictly monotonically decreasing in* σ*, i.e.,*

$$\langle f, \Lambda_\sigma f\rangle_{L^2(T)} > \langle f, \Lambda_{\tilde\sigma} f\rangle_{L^2(T)} \qquad \text{for } \sigma \le \tilde\sigma, \sigma \ne \tilde\sigma, \text{ and } f \ne 0.$$

Proof. (a) Compactness is a consequence of the compactness of the embeddings $H_\diamond^{1/2}(T) \hookrightarrow L_\diamond^2(T)$ and $L_\diamond^2(T) \hookrightarrow H_\diamond^{-1/2}(T)$, and self-adjointness and positivity follow from the weak formulation of (2.1) since

$$\langle f, \Lambda_\sigma \tilde f\rangle_{L^2(T)} = \int_B \sigma \nabla u_\sigma \cdot \nabla \tilde u_\sigma \, dx = \langle \tilde f, \Lambda_\sigma f\rangle_{L^2(T)},$$

where u_σ and $\tilde u_\sigma$ are the solutions of (2.1) for f and $\tilde f$, respectively.

(b) Here we use the fact that u_σ is the unique minimizer in $H_\diamond^1(B)$ of the energy functional

$$\frac{1}{2}\int_B \sigma |\nabla u|^2 \, dx - \oint_T f u \, ds$$

with the minimum value $-\frac{1}{2}\oint_T f u_\sigma \, ds$. Therefore,

$$\begin{aligned} -\frac{1}{2}\langle f, \Lambda_\sigma f\rangle_{L^2(T)} &= \frac{1}{2}\int_B \sigma |\nabla u_\sigma|^2 \, dx - \oint_T f u_\sigma \, ds \\ &< \frac{1}{2}\int_B \sigma |\nabla u_{\tilde\sigma}|^2 \, dx - \oint_T f u_{\tilde\sigma} \, ds \\ &\le \frac{1}{2}\int_B \tilde\sigma |\nabla u_{\tilde\sigma}|^2 \, dx - \oint_T f u_{\tilde\sigma} \, ds = -\frac{1}{2}\langle f, \Lambda_{\tilde\sigma} f\rangle_{L^2(T)}, \end{aligned}$$

which establishes the monotonicity. ∎

For piecewise smooth conductivities of the form

$$\sigma(x) = \begin{cases} \kappa(x), & x \in \overline{\Omega}, \\ 1, & x \in B \setminus \overline{\Omega}, \end{cases} \tag{2.2}$$

the forward problem (2.1) can be formulated in a classical sense as a *diffraction problem* (cf. [16]);

$$\begin{aligned} \nabla \cdot \sigma \nabla u = 0 \quad \text{in } B \setminus \Gamma, \qquad & \frac{\partial u}{\partial \nu} = f \quad \text{on } T, \\ [u]_\Gamma = 0, \qquad & \Big[\sigma \frac{\partial u}{\partial \nu}\Big]_\Gamma = 0. \end{aligned} \tag{2.3}$$

Here, $[\,\cdot\,]_\Gamma$ denotes the jump of the bracketed quantity across the inner boundary $\Gamma = \partial\Omega$, i.e.,

$$[u]_\Gamma = u^+ - u^- \qquad \text{and} \qquad \Big[\sigma \frac{\partial u}{\partial \nu}\Big]_\Gamma = \frac{\partial u^+}{\partial \nu} - \kappa \frac{\partial u^-}{\partial \nu}$$

with

$$u^{\pm}(x) = \lim_{t\to 0+} u(x \pm t\nu) \qquad \text{and} \qquad \frac{\partial u^{\pm}(x)}{\partial \nu} = \lim_{t\to 0+} \nu \cdot \nabla u(x \pm t\nu)$$

for $x \in \Gamma$, where ν is the normal vector at x pointing outward Ω.

Example 2.1 (Radially symmetric case). For the unit disk $B = B_1 = \{x \in \mathbb{R}^2 : |x| < 1\}$ consider the conductivity distribution

$$\sigma(x) = \begin{cases} \kappa, & |x| \le \rho, \\ 1, & \rho < |x| < 1, \end{cases}$$

with a constant $\kappa > 0$ and $0 < \rho < 1$. Solving the forward problem explicitly using polar coordinates $x = re^{i\xi}$ yields the spectral decomposition

$$\Lambda_\sigma : \begin{Bmatrix} \frac{1}{\sqrt{\pi}} \cos k\xi \\ \frac{1}{\sqrt{\pi}} \sin k\xi \end{Bmatrix} \longmapsto \frac{1}{k} \frac{1+\mu\rho^{2k}}{1-\mu\rho^{2k}} \begin{Bmatrix} \frac{1}{\sqrt{\pi}} \cos k\xi \\ \frac{1}{\sqrt{\pi}} \sin k\xi \end{Bmatrix} \tag{2.4}$$

with $\mu = (1-\kappa)/(1+\kappa) \in (-1,1)$; cf., for instance, [18]. The asymptotic decay $\mathcal{O}(k^{-1})$ of the eigenvalues indicates the smoothing, $H_\diamond^{-1/2}(T) \to H_\diamond^{1/2}(T)$, of exactly one order in Sobolev scale of this Neumann–Dirichlet map. Later on this example will be reconsidered in order to illustrate some of our abstract results.

3. Main result (for piecewise constant conductivity). In this section we confine ourselves to the special case of a single inclusion, in which the conductivity is constant, i.e., $\kappa(x) = \kappa$ in (2.2). As will be shown in section 4, our results can be generalized to the case of multiple inclusions and less restrictive conductivity distributions.

Let $N(z,x)$ be a Neumann function for the domain B, i.e., $N(z,x) = \Phi(z-x) + n(z,x)$ for $z, x \in B$, where Φ denotes the fundamental solution for the Laplacian, and $n(z,x)$ solves the Neumann boundary value problem

$$\begin{aligned} \Delta_x n(z,x) &= 0 \quad \text{in } B, \\ \frac{\partial n(z,x)}{\partial \nu_x} &= \beta(x) - \frac{\partial \Phi(z-x)}{\partial \nu_x} \quad \text{on } T, \\ \oint_T \beta(x) n(z,x)\, ds_x &= -\oint_T \beta(x)\Phi(z-x)\, ds_x \end{aligned}$$

for some scaling function β with $\oint_T \beta(x)\, ds = 1$. In our main result we will need the boundary values of the directional derivative of $N(z,x)$ with respect to z in some fixed direction d, $|d| = 1$,

$$g_{z,d}(x) = \frac{\partial N(z,x)}{\partial_z d}\Big|_T = (d \cdot \nabla_z N(z,x))\big|_T.$$

THEOREM 3.1. *For $\kappa \neq 1$ there holds $g_{z,d} \in \mathcal{R}(|\Lambda_\sigma - \Lambda_{\mathbb{1}}|^{1/2})$ if and only if $z \in \Omega$.*

Outline of proof. The key ingredient of the proof is a factorization $\Lambda_\sigma - \Lambda_{\mathbb{1}} = LFL'$, where $L : H_\diamond^{-1/2}(\Gamma) \to H_\diamond^{1/2}(T)$ and $L' : H_\diamond^{-1/2}(T) \to H_\diamond^{1/2}(\Gamma)$ are dual to each other and $F : H_\diamond^{1/2}(\Gamma) \to H_\diamond^{-1/2}(\Gamma)$ is an isomorphism with $F = F'$. This factorization will be derived in section 3.1.

Having this factorization at our disposal, we shall be able to show equality of the ranges, $\mathcal{R}(|\Lambda_\sigma - \Lambda_{\mathbb{1}}|^{1/2}) = \mathcal{R}(L)$. Here, $\mathcal{R}(L) \subset H_\diamond^{1/2}(T)$ is understood as a subspace of $L_\diamond^2(T)$.

Then, the special choice of L allows for a characterization of $\mathcal{R}(L)$ according to the theorem; see section 3.2 for details. ∎

Note that, as a byproduct, Theorem 3.1 also provides a uniqueness result for this class of conductivities and, in contrast to the uniqueness results cited above (apart from [10]), it is proven in a constructive way.

3.1. Factorization of the Neumann–Dirichlet map. Consider the following Neumann boundary value problem:

$$\Delta v = 0 \quad \text{in } B \setminus \overline{\Omega}, \qquad \frac{\partial v}{\partial \nu} = 0 \quad \text{on } T, \qquad \frac{\partial v^+}{\partial \nu} = \phi \quad \text{on } \Gamma, \tag{3.1}$$

which for $\phi \in H_\diamond^{-1/2}(\Gamma)$ has a unique solution $v \in H^1_{\diamond,T}(B \setminus \overline{\Omega})$. Thus, we may define a bounded operator L by

$$L : H_\diamond^{-1/2}(\Gamma) \longrightarrow H_\diamond^{1/2}(T), \qquad \phi \longmapsto v|_T. \tag{3.2}$$

The dual operator L' of L is then given by the boundary value problem

$$\Delta v' = 0 \quad \text{in } B \setminus \overline{\Omega}, \qquad \frac{\partial v'}{\partial \nu} = -\phi' \quad \text{on } T, \qquad \frac{\partial {v'}^+}{\partial \nu} = 0 \quad \text{on } \Gamma, \tag{3.3}$$

with a unique solution $v' \in H^1_{\diamond,\Gamma}(B \setminus \overline{\Omega})$ via

$$L' : H_\diamond^{-1/2}(T) \longrightarrow H_\diamond^{1/2}(\Gamma), \qquad \phi' \longmapsto v'|_\Gamma. \tag{3.4}$$

This is easily seen by applying Green's formula to v and v' in the domain $B \setminus \overline{\Omega}$:

$$\begin{aligned}\langle \phi', L\phi \rangle_{L^2(T)} &= -\oint_T \frac{\partial v'}{\partial \nu} v \, ds \;=\; -\oint_\Gamma \frac{\partial {v'}^+}{\partial \nu} v \, ds - \oint_T v' \frac{\partial v}{\partial \nu} \, ds + \oint_\Gamma v' \frac{\partial v^+}{\partial \nu} \, ds \\ &= \oint_\Gamma v' \phi \, ds \;=\; \langle v', \phi \rangle_{L^2(\Gamma)}.\end{aligned}$$

As follows from uniqueness for the Cauchy problem for the Laplace equation, both L and L' are injective. Note that L and L' depend only on the inclusion Ω itself and not on the specific conductivity distribution in the interior of Ω.

Now we are prepared to formulate our result on the factorization of $\Lambda_\sigma - \Lambda_{\mathbb{1}}$.

LEMMA 3.2. *Let $\kappa \neq 1$. With L and L' defined by* (3.2) *and* (3.4)*, respectively, the difference of Neumann–Dirichlet maps can be factorized as $\Lambda_\sigma - \Lambda_{\mathbb{1}} = LFL'$, where $F : H_\diamond^{1/2}(\Gamma) \to H_\diamond^{-1/2}(\Gamma)$ is an isomorphism with $F = F'$.*

Proof. For fixed boundary current $f \in H_\diamond^{-1/2}(T)$ denote by u_σ the solution of the forward problem (2.3) and by $u_{\mathbb{1}}$ the solution of the associated homogeneous problem. The difference $u_\sigma - u_{\mathbb{1}}$ is then harmonic in $B \setminus \overline{\Omega}$ and

$$\oint_\Gamma \frac{\partial (u_\sigma - u_{\mathbb{1}})^+}{\partial \nu} \, ds = \oint_T \frac{\partial (u_\sigma - u_{\mathbb{1}})}{\partial \nu} \, ds = 0$$

by the divergence theorem. Thus, $v = u_\sigma - u_{\mathbb{1}}$ solves (3.1) for $\phi = \frac{\partial (u_\sigma - u_{\mathbb{1}})^+}{\partial \nu}\big|_\Gamma$ and by the definition (3.2) of L we have

$$L\big(\tfrac{\partial (u_\sigma - u_{\mathbb{1}})^+}{\partial \nu}\big|_\Gamma\big) = (u_\sigma - u_{\mathbb{1}})|_T = (\Lambda_\sigma - \Lambda_{\mathbb{1}}) f. \tag{3.5}$$

If we now introduce the operator $G_\sigma : f \mapsto \frac{\partial u_\sigma^+}{\partial \nu}\big|_\Gamma$ and set $G = G_\sigma - G_{\mathbb{1}}$, then we have so far derived a factorization

$$\Lambda_\sigma - \Lambda_{\mathbb{1}} = LG. \tag{3.6}$$

Note that G_σ is a well-defined bounded operator from $H_\diamond^{-1/2}(T)$ to $H_\diamond^{-1/2}(\Gamma)$. This is a consequence of the boundedness of

$$H(\mathrm{div}, B \setminus \overline{\Omega}) \longrightarrow H^{-1/2}(\Gamma), \qquad \mathbf{v} \longmapsto (\nu \cdot \mathbf{v})^+|_\Gamma, \tag{3.7}$$

the mapping of vector fields in $H(\mathrm{div}, B\setminus\overline{\Omega}) = \{\mathbf{v} \in (L^2(B\setminus\overline{\Omega}))^n : \nabla\cdot\mathbf{v} \in L^2(B\setminus\overline{\Omega})\}$ to their normal component on the boundary (cf. [7, Thm. 2.5]); we apply this result to the current field $\sigma\nabla u_\sigma$.

Next, we want to compute the dual operator G'_σ of G_σ. To this purpose, consider the following diffraction problem with inhomogeneous jump condition:

$$\begin{aligned} \Delta w = 0 \quad &\text{in } B\setminus\Gamma, & \frac{\partial w}{\partial \nu} = 0 \quad &\text{on } T, \\ \big[w\big]_\Gamma = \psi, & & \Big[\sigma\frac{\partial w}{\partial \nu}\Big]_\Gamma = 0, & \end{aligned} \tag{3.8}$$

which for $\psi \in H_\diamond^{1/2}(\Gamma)$ possesses a unique solution w_σ with $w_\sigma|_{B\setminus\overline{\Omega}} \in H^1_{\diamond,T}(B\setminus\overline{\Omega})$ and $w_\sigma|_\Omega \in H^1(\Omega)$; cf., for example, [16]. Applying Green's formula in $B\setminus\overline{\Omega}$ and in Ω, we obtain

$$\begin{aligned} \langle G_\sigma f, \psi\rangle_{L^2(\Gamma)} &= \oint_\Gamma \frac{\partial u_\sigma^+}{\partial\nu}(w_\sigma^+ - w_\sigma^-)\,ds \;=\; -\oint_\Gamma \kappa\frac{\partial u_\sigma^-}{\partial\nu}w_\sigma^-\,ds + \oint_\Gamma \frac{\partial u_\sigma^+}{\partial\nu}w_\sigma^+\,ds \\ &= \oint_\Gamma \Big(-\kappa\frac{\partial w_\sigma^-}{\partial\nu}u_\sigma^- + \frac{\partial w_\sigma^+}{\partial\nu}u_\sigma^+\Big)\,ds + \oint_T \Big(\frac{\partial u_\sigma}{\partial\nu}w_\sigma - \frac{\partial w_\sigma}{\partial\nu}u_\sigma\Big)\,ds \\ &= \oint_T f w_\sigma\,ds \;=\; \langle f, w_\sigma\rangle_{L^2(T)}, \end{aligned}$$

where the first integral in the second line vanishes according to the jump conditions $[u_\sigma]_\Gamma = 0$ and $\big[\sigma\frac{\partial w_\sigma}{\partial\nu}\big]_\Gamma = 0$. This shows that $G'_\sigma\psi = w_\sigma|_T$ and hence

$$G'\psi = (G'_\sigma - G'_{\mathbb{1}})\psi = (w_\sigma - w_{\mathbb{1}})|_T.$$

Note that the restriction of w_σ to $B\setminus\overline{\Omega}$ also solves (3.1) with $\phi = \frac{\partial w_\sigma^+}{\partial\nu}\big|_\Gamma$, which in terms of our operators can be written as $L\big(\frac{\partial w_\sigma^+}{\partial\nu}\big|_\Gamma\big) = w_\sigma|_T = G'_\sigma\psi$, and, by linearity,

$$L\big(\tfrac{\partial (w_\sigma - w_{\mathbb{1}})^+}{\partial\nu}\big|_\Gamma\big) = (w_\sigma - w_{\mathbb{1}})|_T = G'\psi. \tag{3.9}$$

Finally, we define the operator F by the mapping rule $\psi \mapsto \frac{\partial (w_\sigma - w_{\mathbb{1}})^+}{\partial\nu}\big|_\Gamma$. The proof of the asserted properties of the operator F will be deferred to the subsequent Lemma 3.3. With help of this operator (3.9) now reads $LF = G'$, and by transposition we obtain $G = F'L' = FL'$. Inserting this into (3.6) yields the desired factorization $\Lambda_\sigma - \Lambda_{\mathbb{1}} = LFL'$. ∎

LEMMA 3.3. *For $\kappa \neq 1$ the operator $F : \psi \mapsto \frac{\partial (w_\sigma - w_{\mathbb{1}})^+}{\partial\nu}\big|_\Gamma$ is an isomorphism $H_\diamond^{1/2}(\Gamma) \to H_\diamond^{-1/2}(\Gamma)$ with $F' = F$. Here w_σ and $w_{\mathbb{1}}$ are the solutions of the diffraction problem* (3.8) *with conductivities σ and $\mathbb{1}$, respectively.*

Proof. From the solution theory for diffraction problems (see [16]) and (3.7) it follows that $F : H_\diamond^{1/2}(\Gamma) \to H^{-1/2}(\Gamma)$ is well defined and bounded. Moreover, the divergence theorem applied to w_σ in $B \setminus \overline{\Omega}$ gives $\oint_\Gamma \frac{\partial w_\sigma^+}{\partial \nu}\, ds = \oint_T \frac{\partial w_\sigma}{\partial \nu}\, ds = 0$ and similarly $\oint_\Gamma \frac{\partial w_{\mathbb{1}}^+}{\partial \nu}\, ds = 0$, so that we have indeed $F\psi \in H_\diamond^{-1/2}(\Gamma)$.

In order to demonstrate the surjectivity of F we construct for arbitrarily chosen $\phi \in H_\diamond^{-1/2}(\Gamma)$ a function $\psi \in H_\diamond^{1/2}(\Gamma)$ with $F\psi = \phi$. The construction proceeds in several steps. First we take the unique solution $W \in H^1_{\diamond,T}(B \setminus \overline{\Omega})$ of the boundary value problem

$$\Delta W = 0 \quad \text{in } B \setminus \overline{\Omega}, \qquad \frac{\partial W}{\partial \nu} = 0 \quad \text{on } T, \qquad \frac{\partial W^+}{\partial \nu} = \phi \quad \text{on } \Gamma, \tag{3.10}$$

and set $\omega = W^+|_\Gamma$. In the inclusion Ω we define W to be the solution of the Dirichlet problem

$$\Delta W = 0 \quad \text{in } \Omega, \qquad W^- = \omega \quad \text{on } \Gamma, \tag{3.11}$$

and afterwards we set $\varphi = \big[\sigma \frac{\partial W}{\partial \nu}\big]_\Gamma = \phi - \kappa \frac{\partial W^-}{\partial \nu}\big|_\Gamma$; note that the Neumann boundary values $\kappa \frac{\partial W^-}{\partial \nu}\big|_\Gamma$ are well defined in $H^{-1/2}(\Gamma)$ by (3.7) with $B \setminus \overline{\Omega}$ replaced by Ω. Therefore, we have $\varphi \in H_\diamond^{-1/2}(\Gamma)$, because $\oint_\Gamma \frac{\partial W^-}{\partial \nu}\, ds = 0$ by the divergence theorem.

In the next step we define the function $w_{\mathbb{1}}$. In the exterior domain we take $w_{\mathbb{1}}$ to be the unique solution in $H^1_{\diamond,T}(B \setminus \overline{\Omega})$ of the Neumann boundary value problem

$$\Delta w_{\mathbb{1}} = 0 \quad \text{in } B \setminus \overline{\Omega}, \qquad \frac{\partial w_{\mathbb{1}}}{\partial \nu} = 0 \quad \text{on } T, \qquad \frac{\partial w_{\mathbb{1}}^+}{\partial \nu} = \frac{1}{\kappa - 1}\varphi \quad \text{on } \Gamma,$$

whereas in the interior domain we let $w_{\mathbb{1}}$ be the solution of

$$\Delta w_{\mathbb{1}} = 0 \quad \text{in } \Omega, \qquad \frac{\partial w_{\mathbb{1}}^-}{\partial \nu} = \frac{1}{\kappa - 1}\varphi \quad \text{on } \Gamma, \qquad \oint_\Gamma w_{\mathbb{1}}^-\, ds = \oint_\Gamma w_{\mathbb{1}}^+\, ds.$$

Now set $\psi = [w_{\mathbb{1}}]_\Gamma$. The special normalization condition for the last interior problem ensures that $\oint_\Gamma \psi\, ds = 0$ and hence $\psi \in H_\diamond^{1/2}(\Gamma)$.

Then it is easily checked that the so-defined functions $w_{\mathbb{1}}$ and $w_\sigma := w_{\mathbb{1}} + W$ solve the diffraction problems (3.8) for conductivities $\mathbb{1}$ and σ, respectively. Obviously we have $F\psi = \frac{\partial (w_\sigma - w_{\mathbb{1}})^+}{\partial \nu}\big|_\Gamma = \frac{\partial W^+}{\partial \nu}\big|_\Gamma = \phi$, what we intended to show.

To verify the injectivity of F let us assume that $F\psi = 0$. This means that the difference $W = w_\sigma - w_{\mathbb{1}}$ of the solutions of the diffraction problems (3.8) for σ and $\mathbb{1}$, respectively, is itself a solution of (3.10) and (3.11) for $\phi = 0$. This implies $W = 0$, i.e., $w_\sigma = w_{\mathbb{1}}$ in $B \setminus \overline{\Omega}$ and Ω. Combining this and $\big[\frac{\partial w_{\mathbb{1}}}{\partial \nu}\big]_\Gamma = \big[\sigma \frac{\partial w_\sigma}{\partial \nu}\big]_\Gamma = 0$ we obtain

$$\frac{\partial w_\sigma^-}{\partial \nu}\big|_\Gamma = \frac{\partial w_{\mathbb{1}}^-}{\partial \nu}\big|_\Gamma = \frac{\partial w_{\mathbb{1}}^+}{\partial \nu}\big|_\Gamma = \frac{\partial w_\sigma^+}{\partial \nu}\big|_\Gamma = \kappa \frac{\partial w_\sigma^-}{\partial \nu}\big|_\Gamma,$$

from which we conclude that all normal derivatives in this equation must be zero, and hence $w_\sigma = w_{\mathbb{1}}$ must be equal to constants in Ω and as well in $B \setminus \overline{\Omega}$. Thus, $\psi = [w_{\mathbb{1}}]_\Gamma = \text{const}$ and after all the normalization condition $\oint_\Gamma \psi\, ds = 0$ implies $\psi = 0$, which shows the injectivity of F.

As a consequence of the open mapping theorem the inverse of the bijective bounded linear operator $F : H_\diamond^{1/2}(\Gamma) \to H_\diamond^{-1/2}(\Gamma)$ is also bounded.

To establish the remaining assertion $F = F'$, it suffices to show that $F_\sigma = F'_\sigma$ (which encloses $F_{\mathbb{1}} = F'_{\mathbb{1}}$ as a special case). For $\psi_1, \psi_2 \in H^{1/2}_\diamond(\Gamma)$ let w_1, w_2 be the corresponding solutions of the diffraction problem (3.8). Green's formula applied to $B \setminus \overline{\Omega}$ and Ω yields

$$\begin{aligned}
\langle F_\sigma \psi_1, \psi_2 \rangle_{L^2(\Gamma)} &= \oint_\Gamma \frac{\partial w_1^+}{\partial \nu} \psi_2 \, ds \;=\; \oint_\Gamma \frac{\partial w_1^+}{\partial \nu} w_2^+ \, ds - \oint_\Gamma \kappa \frac{\partial w_1^-}{\partial \nu} w_2^- \, ds \\
&= \oint_T \Big(\frac{\partial w_1}{\partial \nu} w_2 - w_1 \frac{\partial w_2}{\partial \nu} \Big) \, ds + \oint_\Gamma \Big(w_1^+ \frac{\partial w_2^+}{\partial \nu} - w_1^- \kappa \frac{\partial w_2^-}{\partial \nu} \Big) \, ds \\
&= \oint_\Gamma (w_1^+ - w_1^-) \frac{\partial w_2^+}{\partial \nu} \, ds \;=\; \oint_\Gamma \psi_1 \frac{\partial w_2^+}{\partial \nu} \, ds \\
&= \langle F_\sigma \psi_2, \psi_1 \rangle_{L^2(\Gamma)},
\end{aligned}$$

hence F_σ is symmetric in the dual pairing $\langle H^{-1/2}_\diamond(\Gamma), H^{1/2}_\diamond(\Gamma) \rangle$, and therefore $F_\sigma = F'_\sigma$. ∎

The proof of Lemma 3.2 is now complete. Before making further use of this result, we take up again our radially symmetric example from above.

Example 3.1. For the situation in Example 2.1 we give explicit representations of the operators occurring in our factorization. Solving (3.1) by separation of variables we compute the following singular value decomposition of L:

$$L : \begin{Bmatrix} \frac{1}{\sqrt{\pi\rho}} \cos k\xi \\ \frac{1}{\sqrt{\pi\rho}} \sin k\xi \end{Bmatrix} \longmapsto \frac{1}{k} \frac{2\rho^{k+1/2}}{1-\rho^{2k}} \begin{Bmatrix} -\frac{1}{\sqrt{\pi}} \cos k\xi \\ -\frac{1}{\sqrt{\pi}} \sin k\xi \end{Bmatrix}.$$

Hence, a singular value decomposition of L' is given by

$$L' : \begin{Bmatrix} \frac{1}{\sqrt{\pi}} \cos k\xi \\ \frac{1}{\sqrt{\pi}} \sin k\xi \end{Bmatrix} \longmapsto \frac{1}{k} \frac{2\rho^{k+1/2}}{1-\rho^{2k}} \begin{Bmatrix} -\frac{1}{\sqrt{\pi\rho}} \cos k\xi \\ -\frac{1}{\sqrt{\pi\rho}} \sin k\xi \end{Bmatrix},$$

which we would have also obtained by solving (3.3) explicitly.

Analogously we derive spectral decompositions of the operators F_σ,

$$F_\sigma : \begin{Bmatrix} \frac{1}{\sqrt{\pi\rho}} \cos k\xi \\ \frac{1}{\sqrt{\pi\rho}} \sin k\xi \end{Bmatrix} \longmapsto \frac{k}{2\rho} \frac{(1-\mu)(\rho^{2k}-1)}{1-\mu\rho^{2k}} \begin{Bmatrix} \frac{1}{\sqrt{\pi\rho}} \cos k\xi \\ \frac{1}{\sqrt{\pi\rho}} \sin k\xi \end{Bmatrix},$$

and $F = F_\sigma - F_{\mathbb{1}}$,

$$F : \begin{Bmatrix} \frac{1}{\sqrt{\pi\rho}} \cos k\xi \\ \frac{1}{\sqrt{\pi\rho}} \sin k\xi \end{Bmatrix} \longmapsto \frac{k}{2\rho} \frac{\mu(\rho^{2k}-1)^2}{1-\mu\rho^{2k}} \begin{Bmatrix} \frac{1}{\sqrt{\pi\rho}} \cos k\xi \\ \frac{1}{\sqrt{\pi\rho}} \sin k\xi \end{Bmatrix}.$$

Here, the asymptotic behavior $\mathcal{O}(k)$ of the eigenvalues parallels the fact that $F : H^{1/2}_\diamond(\Gamma) \to H^{-1/2}_\diamond(\Gamma)$ is an isomorphism.

3.2. Range of L. The next building block for the proof of Theorem 3.1 is to show that the ranges of L and $|\Lambda_\sigma - \Lambda_{\mathbb{1}}|^{1/2}$ coincide.

We set $\mathcal{D}(F) = F^{-1}(L^2_\diamond(\Gamma))$ and interpret F as an operator from $\mathcal{D}(F)$ onto $L^2_\diamond(\Gamma)$; note that $\mathcal{D}(F)$ is a proper subspace of $H^{1/2}_\diamond(\Gamma)$ by Lemma 3.3. Since $F = F'$, the restriction $F^{-1} : L^2_\diamond(\Gamma) \to \mathcal{D}(F)$ is self-adjoint, and as the inverse of an injective self-adjoint operator is itself self-adjoint (see, e.g., [19, Thm. 13.11]), we obtain that $F : \mathcal{D}(F) \to L^2_\diamond(\Gamma)$ is self-adjoint; in particular, F is densely defined in $L^2_\diamond(\Gamma)$.

Next we show that $\pm F$ is positive, where the sign is chosen according to the sign of $1-\kappa$. For $0 \neq \psi \in \mathcal{R}(L') \subset H_\diamond^{1/2}(\Gamma)$, i.e., $\psi = L'\phi$ with $\phi \in H_\diamond^{-1/2}(T)$, we have

$$\begin{aligned}\langle F\psi, \psi\rangle_{L^2(\Gamma)} &= \langle FL'\phi, L'\phi\rangle_{L^2(\Gamma)} = \langle LFL'\phi, \phi\rangle_{L^2(T)} \\ &= \langle (\Lambda_\sigma - \Lambda_{\mathbb{1}})\phi, \phi\rangle_{L^2(T)} \begin{cases} > 0 & \text{for } 0 < \kappa < 1, \\ < 0 & \text{for } \kappa > 1, \end{cases}\end{aligned}$$

using the factorization of $\Lambda_\sigma - \Lambda_{\mathbb{1}}$ and Lemma 2.1(b). But since L is injective and therefore $\overline{\mathcal{R}(L')} = \mathcal{N}(L)^\perp = \{0\}^\perp = H_\diamond^{1/2}(\Gamma)$, the range of L' is dense in $H_\diamond^{1/2}(\Gamma)$; thus we conclude that for all $\psi \in H_\diamond^{1/2}(\Gamma)$

$$\langle F\psi, \psi\rangle_{L^2(\Gamma)} \begin{cases} \geq 0 & \text{for } 0 < \kappa < 1, \\ \leq 0 & \text{for } \kappa > 1. \end{cases}$$

Since F is injective and densely defined, the case $\langle \psi, F\psi\rangle_{L^2(\Gamma)} = 0$ for $0 \neq \psi \in \mathcal{D}(F)$ can be excluded, whence the positivity of $\pm F$ follows.[1]

We proceed with a functional analytic auxiliary result.

LEMMA 3.4. *Let $V \hookrightarrow H \hookrightarrow V'$ be a Gelfand triple, i.e., these injections are continuous and have dense range. Assume that the restriction of an isomorphism $K : V' \to V$ to an operator $H \to H$ is self-adjoint and positive. Then, the (self-adjoint and positive) square root $K^{1/2} : H \to H$ is an isomorphism $H \to V$ and admits an isomorphic extension $\widetilde{K}^{1/2} : V' \to H$. Furthermore, these two isomorphisms are dual to each other, and $K : V' \to V$ can be written as $K = K^{1/2}\widetilde{K}^{1/2}$.*

Proof. Since we have

$$\|K^{1/2}\phi\|_H^2 = \langle K\phi, \phi\rangle_H \leq \|K\phi\|_V \|\phi\|_{V'} \leq \|K\|_{V'\to V}\|\phi\|_{V'}^2 \qquad \text{for } \phi \in H,$$

the square root $K^{1/2}$ can be extended to a bounded operator $\widetilde{K}^{1/2} : V' \to H$. By duality, $(\widetilde{K}^{1/2})' : H \to V$ is also bounded and coincides on H with $\widetilde{K}^{1/2}|_H = K^{1/2}$ because of the self-adjointness. Hence, $K^{1/2}$ is bounded as an operator $H \to V$ and moreover we have

$$K = K^{1/2}\widetilde{K}^{1/2} : V' \to V, \tag{3.12}$$

for the operators on both sides are bounded and coincide on the dense subspace $H \subset V'$.

The operator $K^{1/2}$ is injective as a square root of an injective operator, and (3.12) and the bijectivity of K imply also its surjectivity. Finally, the bijectivity of $\widetilde{K}^{1/2}$ is now again a consequence of (3.12). ∎

Applying this lemma to $K = |F|^{-1}$ and taking inverses, we see that the square root of $|F|$ can be extended to isomorphic operators $|F|^{1/2} : H_\diamond^{1/2}(\Gamma) \to L_\diamond^2(\Gamma)$ and $(|F|^{1/2})' : L_\diamond^2(\Gamma) \to H_\diamond^{-1/2}(\Gamma)$, so that $|F| : H_\diamond^{1/2}(\Gamma) \to H_\diamond^{-1/2}(\Gamma)$ can be written as $|F| = (|F|^{1/2})'|F|^{1/2}$. The factorization derived in Lemma 3.2 of $\Lambda_\sigma - \Lambda_{\mathbb{1}}$, treated as an operator in $L_\diamond^2(T)$, then takes the form

$$|\Lambda_\sigma - \Lambda_{\mathbb{1}}| = L(|F|^{1/2})'|F|^{1/2}L'\big|_{L_\diamond^2(T)} = \big(|F|^{1/2}L'\big|_{L_\diamond^2(T)}\big)^*|F|^{1/2}L'\big|_{L_\diamond^2(T)},$$

where the star denotes the adjoint of $|F|^{1/2}L'\big|_{L_\diamond^2(T)} : L_\diamond^2(T) \to L_\diamond^2(\Gamma)$.

[1] In what follows we write $|F|$ for $\operatorname{sgn}(1-\kappa)F$.

Now, we can apply a general result to our case, namely, that for every operator A acting between Hilbert spaces there holds $\mathcal{R}((A^*A)^{1/2}) = \mathcal{R}(A^*)$; see, e.g., [6, Prop. 2.18]. This yields

$$\mathcal{R}(|\Lambda_\sigma - \Lambda_{\mathbb{1}}|^{1/2}) = \mathcal{R}((|F|^{1/2}L'\big|_{L^2_\diamond(T)})^*) = \mathcal{R}(L(|F|^{1/2})') = \mathcal{R}(L), \tag{3.13}$$

the latter equality due to the bijectivity of $(|F|^{1/2})' : L^2_\diamond(\Gamma) \to H_\diamond^{-1/2}(\Gamma)$. This is just what we wanted to show.

To complete the proof of Theorem 3.1 we have to figure out the connection between the boundary values $g_{z,d} = d \cdot \nabla_z N(z, \cdot\,)\big|_T$ and the range of L.

LEMMA 3.5. *$g_{z,d} \in \mathcal{R}(L)$ if and only if $z \in \Omega$.*

Proof. It is obvious from (3.1) and (3.2) that $g_{z,d}$ is contained in the range of the operator L if and only if the Cauchy problem

$$\begin{aligned} \Delta v &= 0 \quad \text{in } B \setminus \overline{\Omega}, \\ v &= g_{z,d} \quad \text{on } T, \qquad \frac{\partial v}{\partial \nu} = 0 \quad \text{on } T, \end{aligned} \tag{3.14}$$

has a solution $v \in H^1_{\diamond,T}(B \setminus \overline{\Omega})$. In this case we would have $\frac{\partial v^+}{\partial \nu}\big|_\Gamma \in H_\diamond^{-1/2}(\Gamma)$ and $L(\frac{\partial v^+}{\partial \nu}\big|_\Gamma) = g_{z,d}$.

Let us recall that $d \cdot \nabla_z N(z, \cdot\,)$ is harmonic away from z where this function has a dipole-like singularity. Moreover, for $z \in B$ and $x \in T$ we have

$$\frac{\partial}{\partial \nu_x} d \cdot \nabla_z N(z, \cdot\,) = d \cdot \nabla_z \frac{\partial N(z, \cdot\,)}{\partial \nu_x} = d \cdot \nabla_z \beta(x) = 0$$

and $d \cdot \nabla_z N(z, \cdot\,) = g_{z,d}$; hence $d \cdot \nabla_z N(z, \cdot\,)$ fulfills the Cauchy boundary conditions in (3.14).

From uniqueness for the Cauchy problem for the Laplace equation we can now deduce that $d \cdot \nabla_z N(z, \cdot\,)$ is the only possible candidate for a solution of the Cauchy problem (3.14). If $z \in \Omega$, then indeed $d \cdot \nabla_z N(z, \cdot\,)$ is a solution, whereas if the singularity z lies in $B \setminus \overline{\Omega}$ or on the inner boundary Γ, then $d \cdot \nabla_z N(z, \cdot\,) \notin H^1_{\diamond,T}(B \setminus \overline{\Omega})$, i.e., it is not a solution of (3.14). This proves the lemma. ∎

Example 3.2. The Neumann function for the unit disk and for the scaling function $\beta \equiv 1/2\pi$ is known explicitly, namely,

$$N(z,x) = \begin{cases} \frac{1}{2\pi}\big(\log|z-x| + \log\big|\frac{z}{|z|} - |z|x\big|\big) & \text{for } z \neq 0, \\ \frac{1}{2\pi}\log|x| & \text{for } z = 0, \end{cases}$$

and the boundary values of its directional derivative are

$$g_{z,d}(x) = d \cdot \nabla_z N(z,x)\big|_T = \frac{1}{\pi}\frac{(z-x)\cdot d}{|z-x|^2} \qquad \text{for } |x| = 1.$$

Introducing polar coordinates $z = |z|e^{i\zeta}$, $d = e^{i\vartheta}$, $x = e^{i\xi}$ we obtain for $g_{z,d}$ the Fourier series expansion

$$g_{z,d}(\xi) = \sum_{k=1}^{\infty} \frac{-|z|^{k-1}}{\pi} \cos\big(k(\xi-\zeta) - (\vartheta - \zeta)\big).$$

Then the (formal) solution of the Cauchy problem (3.14) reads

$$v(r,\xi)=\sum_{k=1}^{\infty}\frac{-|z|^{k-1}}{2\pi}\cos\big(k(\xi-\zeta)-(\vartheta-\zeta)\big)\big(r^k+r^{-k}\big),$$

and the normal derivative on the circle $\Gamma=\{z:|z|=\rho\}$ can (formally) be computed as

$$\frac{\partial v}{\partial r}(\rho,\xi)=\sum_{k=1}^{\infty}\frac{-k|z|^{k-1}}{2\pi}\frac{\cos\big(k(\xi-\zeta)-(\vartheta-\zeta)\big)}{\sqrt{\rho}}\big(\rho^{k-1/2}-\rho^{-k-1/2}\big).$$

This function belongs to $H_{\diamond}^{-1/2}(\Gamma)$ if and only if the sum

$$\sum_{k=1}^{\infty}k^{-1}\Big(\frac{-k|z|^{k-1}}{2\pi}\Big)^2\big(\rho^{k-1/2}-\rho^{-k-1/2}\big)^2 = \sum_{k=1}^{\infty}\frac{k|z|^{2k-2}(\rho^{2k-1}-2\rho^{-1})}{4\pi^2}+\frac{1}{4\pi^2\rho^3}\sum_{k=1}^{\infty}k\Big(\frac{|z|}{\rho}\Big)^{2k-2}$$

converges. While the first sum is always finite, for the second sum this is the case if and only if $\rho>|z|$, i.e., if the test point z is located in the interior of the inclusion Ω. This illustrates the evidence of Lemma 3.5.

According to Theorem 3.1 we may alternatively use the spectral decomposition of $\Lambda_\sigma-\Lambda_{\mathbb{1}}$ (cf. (2.4)),

$$\Lambda_\sigma-\Lambda_{\mathbb{1}}:\begin{Bmatrix}\frac{1}{\sqrt{\pi}}\cos k\xi\\ \frac{1}{\sqrt{\pi}}\sin k\xi\end{Bmatrix}\longmapsto\frac{2}{k}\frac{\mu\rho^{2k}}{1-\mu\rho^{2k}}\begin{Bmatrix}\frac{1}{\sqrt{\pi}}\cos k\xi\\ \frac{1}{\sqrt{\pi}}\sin k\xi\end{Bmatrix}.$$

Indeed, we can actually calculate (formally) the preimage of $g_{z,d}$ under $|\Lambda_\sigma-\Lambda_{\mathbb{1}}|^{1/2}$, namely,

$$|\Lambda_\sigma-\Lambda_{\mathbb{1}}|^{-1/2}g_{z,d}=\sum_{k=1}^{\infty}-\frac{|z|^{k-1}}{\pi}\Big|\frac{2}{k}\frac{\mu\rho^{2k}}{1-\mu\rho^{2k}}\Big|^{-1/2}\cos\big(k(\xi-\zeta)-(\vartheta-\zeta)\big),$$

which is an element of $L^2_{\diamond}(T)$ if and only if the Fourier coefficients are square summable. Obviously, this takes place if and only if $|z|<\rho$, thus leading to the same result as predicted by Theorem 3.1.

4. Generalizations.

4.1. Multiple inclusions. Our first generalization concerns the practically important case of finitely many separated inclusions. By this we mean simply connected open domains $\Omega_1,\dots,\Omega_p$ with $\overline{\Omega}_i\cap\overline{\Omega}_j=\emptyset$ for $i\neq j$. Our aim is to carry over Theorem 3.1 to the case of piecewise constant conductivities, i.e.,

$$\sigma(x)=\begin{cases}\kappa, & x\in\overline{\Omega}_j,\quad j=1,\dots,p,\\ 1, & x\in B\setminus\overline{\Omega},\end{cases}\tag{4.1}$$

where $\Omega=\Omega_1\cup\cdots\cup\Omega_p$. Basically the proof for a single inclusion can be adopted with few minor modifications on which we will comment now.

Since the inner boundary now consists of several components $\Gamma_j = \partial\Omega_j$, it is convenient to set $\Gamma = \Gamma_1 \times \cdots \times \Gamma_p$ and interpret the relevant Sobolev spaces accordingly as product spaces, e.g., $H_\diamond^{\pm 1/2}(\Gamma) = H_\diamond^{\pm 1/2}(\Gamma_1) \times \cdots \times H_\diamond^{\pm 1/2}(\Gamma_p)$.

The operator L is again defined by (3.1), (3.2), where the inner Neumann boundary condition should be understood componentwise, i.e., $\frac{\partial v^+}{\partial \nu} = \phi_j$ on Γ_j, $j = 1, \ldots, p$, for $\phi = (\phi_1, \ldots, \phi_p) \in H_\diamond^{-1/2}(\Gamma)$.

For the definition of its dual operator we consider again the boundary value problem (3.3), whose solution v' is unique up to an additive constant. Special care now has to be taken because in general no solution fulfills all normalization conditions $\oint_{\Gamma_j} v'\, ds = 0$, $j = 1, \ldots, p$, at once. It turns out that each trace component $v'|_{\Gamma_j}$ has to be normalized independently, i.e., if we fix an arbitrary solution v' and set $c_j = \big(\oint_{\Gamma_j} v'\, ds\big) \,/\, \big(\oint_{\Gamma_j} ds\big)$, then one verifies that the dual operator of L is given by

$$L' : H_\diamond^{-1/2}(T) \longrightarrow H_\diamond^{1/2}(\Gamma), \qquad \phi' \longmapsto (v'|_{\Gamma_1} - c_1, \ldots, v'|_{\Gamma_p} - c_p).$$

The proofs of the factorization in Lemma 3.2 and the range characterization in section 3.2 then remain essentially unchanged if boundary conditions on Γ are interpreted componentwise. Thus we have shown:

PROPOSITION 4.1. *Theorem* 3.1 *holds also true in the case* (4.1) *of multiple inclusions.*

4.2. Insulating and perfectly conducting inclusions. Next we extend Theorem 3.1 to the extreme cases of insulating ($\kappa = 0$) and perfectly conducting ($\kappa = \infty$) inclusions, for which the forward problems must be modeled slightly differently.

No current can flow into an insulating inclusion; this is expressed by the inner boundary condition $\frac{\partial u^+}{\partial \nu} = 0$ on Γ. Thus, the forward problem can be formulated as follows:

$$\Delta u = 0 \quad \text{in } B \setminus \overline{\Omega}, \qquad \frac{\partial u}{\partial \nu} = f \quad \text{on } T, \qquad \frac{\partial u^+}{\partial \nu} = 0 \quad \text{on } \Gamma,$$

for which a unique weak solution $u_\sigma \in H^1_{\diamond,T}(B \setminus \overline{\Omega})$ exists; here, by a weak solution we mean $u_\sigma \in H^1_{\diamond,T}(B \setminus \overline{\Omega})$ satisfying

$$\int_{B\setminus\overline{\Omega}} \nabla u_\sigma \cdot \nabla \phi \, dx = \oint_T f\phi \, ds \qquad \text{for all } \phi \in H^1_{\diamond,T}(B \setminus \overline{\Omega}).$$

In perfectly conducting inclusions potential differences would even out immediately so that there the potential must be constant. Consequently, also in this case the differential equation needs only to be considered in $B \setminus \overline{\Omega}$ if one encloses the source freedom of the inclusion as an additional condition, which implicitly determines the value of the potential constant inside of Ω. The forward problem thus may be stated as follows:

$$\begin{aligned} &\Delta u = 0 \quad \text{in } B \setminus \overline{\Omega}, && \frac{\partial u}{\partial \nu} = f \quad \text{on } T, \\ &u^+ = \text{const} \quad \text{on } \Gamma, && \oint_\Gamma \frac{\partial u^+}{\partial \nu}\, ds = 0, \end{aligned} \tag{4.2}$$

for which we seek a solution u_σ normalized according to $\oint_T \frac{\partial u^+}{\partial \nu}\, ds = 0$. By means of the subspace

$$\tilde{H}^1_{\diamond,T}(B \setminus \overline{\Omega}) = \{u \in H^1_{\diamond,T}(B \setminus \overline{\Omega}) : u|_\Gamma = \text{const}\}$$

the corresponding weak formulation then reads as follows: determine $u_\sigma \in \tilde{H}^1_{\diamond,T}(B\setminus\overline{\Omega})$ satisfying

$$\int_{B\setminus\overline{\Omega}} \nabla u_\sigma \cdot \nabla\phi \, dx = \oint_T f\phi \, ds \qquad \text{for all } \phi \in \tilde{H}^1_{\diamond,T}(B \setminus \overline{\Omega}).$$

In both the insulating and the perfectly conducting cases the solution u_σ is also given as the minimizer of

$$\frac{1}{2}\int_{B\setminus\overline{\Omega}} |\nabla u|^2 \, dx - \oint_T fu \, ds$$

in the spaces $H^1_{\diamond,T}(B\setminus\overline{\Omega})$ and $\tilde{H}^1_{\diamond,T}(B\setminus\overline{\Omega})$, respectively. This fact allows us to extend the monotonicity property of Lemma 2.1(b) to these two extreme cases.

The factorization of the Neumann–Dirichlet map, $\Lambda_\sigma - \Lambda_{\mathbb{1}} = L'FL$, follows along the lines of section 3.1. Since as stressed above the operators L and L' depend anyhow only on Ω, merely the definition of F has to be adjusted. However, the properties of F shown in Lemma 3.3 also hold in these cases; see [2] for details. The proof of the equality (3.13) relies only on these properties and on the definiteness of $\Lambda_\sigma - \Lambda_{\mathbb{1}}$, which follows from the monotonicity.

PROPOSITION 4.2. *Theorem* 3.1 *is also valid if* $\kappa = 0$ *or* $\kappa = \infty$.

It should be emphasized here that Hähner [9] recently gave a similar characterization for grounded perfectly conducting inclusions. There, the grounding is expressed by zero Dirichlet boundary conditions on the inner boundary Γ replacing the two last conditions in (4.2).

4.3. Nonconstant conductivity inside inclusions. Now we want to generalize Theorem 3.1 to our initial situation (1.2), where the conductivity is allowed to vary within Ω subject to the constraint (1.3). The idea here is to reduce this case to the already proven results by sandwiching the nonconstant conductivity distribution between piecewise constant ones.

Let us assume that $\kappa(x) \geq 1+\varepsilon$ (the case $\kappa(x) \leq 1-\varepsilon$ can be treated analogously). If we set

$$\underline{\sigma}(x) = \begin{cases} 1+\varepsilon, & x \in \Omega, \\ 1, & x \in B\setminus\overline{\Omega}, \end{cases} \qquad \text{and} \qquad \overline{\sigma}(x) = \begin{cases} \infty, & x \in \Omega, \\ 1, & x \in B\setminus\overline{\Omega}, \end{cases}$$

then we have $\underline{\sigma} \leq \sigma \leq \overline{\sigma}$ and by the monotonicity of the Neumann–Dirichlet map

$$\Lambda_{\overline{\sigma}} - \Lambda_{\mathbb{1}} \leq \Lambda_\sigma - \Lambda_{\mathbb{1}} \leq \Lambda_{\underline{\sigma}} - \Lambda_{\mathbb{1}} \leq O, \tag{4.3}$$

where the ordering of the operators is to be understood in the sense of positive semidefiniteness as in Lemma 2.1(b).

Now we utilize an auxiliary result, namely, that for a self-adjoint and injective operator A we have

$$y \in \mathcal{R}(A) \qquad \text{if and only if} \qquad \sup_{0\neq x\in\mathcal{D}(A)} \frac{\langle y, x\rangle}{\|Ax\|} < \infty. \tag{4.4}$$

This is seen as follows: first note that the inverse of an injective self-adjoint operator A is itself self-adjoint and $\mathcal{R}(A) = \mathcal{D}(A^{-1}) = \mathcal{D}(A^{-*})$. But by definition we have

$y \in \mathcal{D}(A^{-*})$ if and only if $\sup\{\frac{\langle A^{-1}z,y\rangle}{\|z\|} : 0 \neq z \in \mathcal{D}(A^{-1})\} < \infty$; see, e.g., [19, sect. 13]. Using $\mathcal{D}(A^{-1}) = \mathcal{R}(A)$ and substituting $z = Ax$ then yields (4.4).

We apply this to our situation by letting $y \in \mathcal{R}(|\Lambda_\sigma - \Lambda_{\mathbb{1}}|^{1/2})$. From (4.3) and (4.4) we obtain

$$\sup_{0\neq x\in L^2_\diamond(T)} \frac{\langle y,x\rangle_{L^2(T)}}{\||\Lambda_{\overline{\sigma}} - \Lambda_{\mathbb{1}}|^{1/2}x\|_{L^2_\diamond(T)}} \leq \sup_{0\neq x\in L^2_\diamond(T)} \frac{\langle y,x\rangle_{L^2(T)}}{\||\Lambda_\sigma - \Lambda_{\mathbb{1}}|^{1/2}x\|_{L^2_\diamond(T)}} < \infty,$$

and then it follows that $y \in \mathcal{R}(|\Lambda_{\overline{\sigma}} - \Lambda_{\mathbb{1}}|^{1/2})$ again by (4.4). Thus we have $\mathcal{R}(|\Lambda_\sigma - \Lambda_{\mathbb{1}}|^{1/2}) \subset \mathcal{R}(|\Lambda_{\overline{\sigma}} - \Lambda_{\mathbb{1}}|^{1/2})$ and correspondingly we can show that $\mathcal{R}(|\Lambda_{\underline{\sigma}} - \Lambda_{\mathbb{1}}|^{1/2}) \subset \mathcal{R}(|\Lambda_\sigma - \Lambda_{\mathbb{1}}|^{1/2})$. But since $\mathcal{R}(|\Lambda_{\underline{\sigma}} - \Lambda_{\mathbb{1}}|^{1/2}) = \mathcal{R}(|\Lambda_{\overline{\sigma}} - \Lambda_{\mathbb{1}}|^{1/2}) = \mathcal{R}(L)$ (cf. (3.13)), $\mathcal{R}(|\Lambda_\sigma - \Lambda_{\mathbb{1}}|^{1/2})$ also must coincide with $\mathcal{R}(L)$.

PROPOSITION 4.3. *Theorem* 3.1 *is also valid for conductivity distributions of the form* (1.2) *subject to the constraint* (1.3).

4.4. Nonconstant and anisotropic background. Our proofs can be adapted in a straightforward way to the case of smoothly varying background conductivities which may even be anisotropic. We emphasize that for our method the background conductivity is required to be a priori known in each of these situations, and only the support of the discontinuous perturbations to this background can be reconstructed.

Let us consider conductivity distributions of the form

$$\sigma(x) = \begin{cases} \kappa(x), & x \in \overline{\Omega}, \\ \gamma(x), & x \in B \setminus \overline{\Omega}, \end{cases} \tag{4.5}$$

where the *background conductivity* $\gamma(x) = (\gamma_{ij}(x))$ is a smooth and symmetric positive definite matrix in $\overline{B}$, and the conductivity inside the inclusion, $\kappa(x) = (\kappa_{ij}(x))$, is symmetric positive definite. The discontinuity condition now takes the form

$$\kappa(x) \geq (1+\varepsilon)\gamma(x) \quad \text{or} \quad \kappa(x) \leq (1-\varepsilon)\gamma(x) \qquad \text{for } x \in \overline{\Omega} \text{ and some } \varepsilon > 0, \tag{4.6}$$

where the matrix ordering is to be understood in the sense of positive semidefiniteness.

As test functions we utilize the boundary values $g_{z,d}(x) = (d \cdot \nabla_z N(z,x))|_T$ of a directional derivative of the Neumann function $N(z,x)$ associated with the differential operator $\nabla\cdot\gamma\nabla$ in B. The Neumann–Dirichlet operator corresponding to γ is denoted by Λ_γ.

PROPOSITION 4.4. *Under the assumption* (4.6) *we have* $g_{z,d} \in \mathcal{R}(|\Lambda_\sigma - \Lambda_\gamma|^{1/2})$ *if and only if* $z \in \Omega$.

Sketch of proof. If $\kappa(x) = \kappa\gamma(x)$ for a constant κ, then one can go step by step through the proof of section 3, replacing the Laplacian by $\nabla\cdot\gamma\nabla$ at every occurrence. The case of more general $\kappa(x)$ is treated by employing the "sandwiching technique" from section 4.3. ∎

4.5. Putting it all together. Propositions 4.1 to 4.4 generalize our findings in several directions independently. However, these extensions can be combined in a straightforward way. For instance, if the conductivity has the form

$$\sigma(x) = \begin{cases} \kappa_j(x), & x \in \overline{\Omega}_j, \quad j = 1,\dots,p, \\ \gamma(x), & x \in B \setminus \overline{\Omega}, \end{cases}$$

of section 4.4 with *several inclusions* Ω_j with $0 \leq \kappa_j(x) \leq (1-\varepsilon)\gamma(x)$ for $x \in \overline{\Omega}_j$, $j = 1,\dots,p$, then one would first sandwich κ in $\overline{\Omega}$ between 0 and $(1-\varepsilon)\gamma(x)$ as in section 4.3 and then proceed as in sections 4.1, 4.2, and 4.4.

Acknowledgements. The author is indebted to Martin Hanke for countless discussions about the subject and for many detailed hints, which improved the presentation of this paper considerably. The author is also grateful to Andreas Kirsch for the continuous readiness to share his ideas for the related inverse scattering problem, and to Michael Pidcock for inspiring and encouraging discussions.

REFERENCES

[1] R. M. Brown and G. A. Uhlmann, *Uniqueness in the inverse conductivity problem for nonsmooth conductivities in two dimensions*, Comm. Partial Differential Equations, 22 (1997), pp. 1009–1027.

[2] M. Brühl, *Gebietserkennung in der elektrischen Impedanztomographie*, Dissertation, Universität Karlsruhe, Karlsruhe, Germany, 1999.

[3] M. Brühl and M. Hanke, *Numerical implementation of two noniterative methods for locating inclusions by impedance tomography*, Inverse Problems, 16 (2000), pp. 1029–1042.

[4] A. P. Calderón, *On an inverse boundary value problem*, in Seminar on Numerical Analysis and its Application to Continuum Physics, W. H. Meyer and M. A. Raupp, eds., Rio de Janeiro, 1980, Soc. Brasil. Mat., Rio de Janeiro, Brazil, 1980, pp. 65–73.

[5] D. Colton and A. Kirsch, *A simple method for solving inverse scattering problems in the resonance region*, Inverse Problems, 12 (1996), pp. 383–393.

[6] H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems*, Kluwer, Dordrecht, The Netherlands, 1996.

[7] V. Girault and P.-A. Raviart, *Finite Element Methods for Navier-Stokes Equations*, Springer, Berlin, 1986.

[8] D. G. Gisser, D. Isaacson, and J. C. Newell, *Electric current computed tomography and eigenvalues*, SIAM J. Appl. Math., 50 (1990), pp. 1623–1634.

[9] P. Hähner, *An inverse problem in electrostatics*, Inverse Problems, 15 (1999), pp. 961–975.

[10] M. Ikehata, *Reconstruction of the support function for inclusion from boundary measurements*, J. Inverse Ill-Posed Probl., 8 (2000), pp. 367–378.

[11] M. Ikehata, *Reconstruction of the shape of the inclusion by boundary measurements*, Comm. Partial Differential Equations, 23 (1998), pp. 1459–1474.

[12] V. Isakov, *On uniqueness of recovery of a discontinuous conductivity coefficient*, Comm. Pure Appl. Math., 41 (1988), pp. 865–877.

[13] A. Kirsch, *Characterization of the shape of the scattering obstacle using the spectral data of the far field operator*, Inverse Problems, 14 (1998), pp. 1489–1512.

[14] R. V. Kohn and M. Vogelius, *Determining conductivity by boundary measurements*, Comm. Pure Appl. Math., 37 (1984), pp. 289–298.

[15] R. V. Kohn and M. Vogelius, *Determining conductivity by boundary measurements* II. *Interior results*, Comm. Pure Appl. Math., 38 (1985), pp. 643–667.

[16] O. A. Ladyzhenskaya, *The Boundary Value Problems of Mathematical Physics*, Springer, New York, 1985.

[17] A. I. Nachman, *Global uniqueness for a two-dimensional inverse boundary value problem*, Ann. of Math. (2), 143 (1996), pp. 71–96.

[18] M. K. Pidcock, M. Kuzuoglu, and K. Leblebicioglu, *Analytic and semi-analytic solutions in electrical impedance tomography:* I. *Two-dimensional problems*, Physiol. Meas., 16 (1995), pp. 77–90.

[19] W. Rudin, *Functional Analysis*, McGraw-Hill, New York, 1973.

[20] J. Sylvester and G. Uhlmann, *A global uniqueness theorem for an inverse boundary value problem*, Ann. of Math. (2), 125 (1987), pp. 153–169.

SIAM J. MATH. ANAL.
Vol. 32, No. 6, pp. 1342–1374

TRAVELLING WAVES FOR FOURTH ORDER PARABOLIC EQUATIONS*

JAN BOUWE VAN DEN BERG†, JOSEPHUS HULSHOF‡, AND ROBERTUS C. VANDERVORST†§

Abstract. We study travelling wave solutions for a class of fourth order parabolic equations. Travelling wave fronts of the form $u(x,t) = U(x+ct)$, connecting homogeneous states, are proven to exist in various cases: connections between two stable states, as well as connections between an unstable and a stable state, are considered.

Key words. travelling wave, fourth order equation, nonlinear parabolic equation, bistable equation, Poincaré transformation, global analysis

AMS subject classifications. 34A34, 34C20, 34C37, 34C23, 35A18, 35K25, 35K55, 37B25, 37C29

PII. S0036141099358300

1. Introduction.

Fourth order parabolic equations of the form

$$u_t = -\gamma u_{xxxx} + u_{xx} + f(u), \qquad \gamma > 0, \tag{1.1}$$

where $x \in \mathbb{R}$, $t > 0$, occur in many physical models such as the theory of phase-transitions [9], nonlinear optics [1], shallow water waves [7], etc. Usually the potential $F(u) = \int f(s)ds$ has at least two local maxima (stable state) and one local minimum (unstable state).[1] A prototypical example is $f_a(u) = (u+a)(1-u^2)$ with $-1 < a < 1$.

For a thorough understanding of (1.1), the stationary problem is of great importance. An extensive literature on this subject exists (see, e.g., [3, 29, 7, 16, 17, 18, 25, 22, 23, 24]). Typically, depending on the parameter γ, the stationary problem displays a multitude of periodic, homoclinic, and heteroclinic solutions. The stationary equation is Hamiltonian, which restricts the possible connections between the equilibrium points. As an example we mention that when the maximum of F is attained in two points, e.g., $F(u) = -\frac{1}{4}(u^2-1)^2$, a solution connecting these maxima exists for all $\gamma > 0$. One could regard this solution as a standing wave. The heteroclinic solution is unique (modulo the obvious symmetries) for small values of γ, say, $\gamma \le \gamma_1(f)$ [28, 29, 19]. On the other hand, for large γ, say, $\gamma > \gamma_2(f)$, there is a multitude of (multibump/transition) solutions connecting the two maxima [17, 18, 24]. This is due to the fact that as γ crosses the critical value $\gamma = \gamma_2(f)$, the eigenvalues of the linearized stationary equation around the two maxima of F become complex.

In the special case $f(u) = u - u^3$, corresponding to $F(u) = -\frac{1}{4}(u^2-1)^2$, it holds that $\gamma_1(f) = \gamma_2(f) = \frac{1}{8}$. Although in many simple cases equality holds, generally

*Received by the editors June 28, 1999; accepted for publication (in revised form) November 1, 2000; published electronically March 28, 2001. This work was partially supported by grant TMR ERBFMRXCT980201.

http://www.siam.org/journals/sima/32-6/35830.html

†Mathematical Institute, Leiden University, P.O. Box 9512, 2300 RA Leiden, The Netherlands (gvdberg@math.leidenuniv.nl, http://www.math.leidenuniv.nl/~gvdberg/; vanderv@math.leidenuniv.nl).

‡Leiden University, Niels Bohrweg L, 2333 CA Leiden, The Netherlands (hulshof@math.leidenuniv.nl, http://www.math.leidenuniv.nl/~hulshof/).

§Center for Dynamical Systems and Nonlinear Studies, Georgia Institute of Technology, Atlanta, GA 30332-0190 (rvander@math.gatech.edu, http://www.math.gatech.edu/ rvander/).

[1]Sometimes the potential is denoted by $-F$ so that the stable states correspond to local minima.

there will be a gap between $\gamma_1(f)$ and $\gamma_2(f)$. The critical value γ_1 is not necessarily small, and a lower bound on γ_1 can in general be explicitly determined (see [29] for more details).

For the time-dependent problem travelling fronts of the form $u(x,t) = U(x + ct)$, connecting extrema of the potential F, play a prominent role in most models. Results on travelling waves for (1.1) have previously been obtained in [6], where nonlinearities of the form $f(u) = f_a(u) = (u + a)(1 - u^2)$, $a \approx 0$, are considered using transversality arguments and perturbing near a standing wave. Moreover, in [2] singular perturbations techniques were applied near $\gamma = 0$. In both cases travelling waves between local maxima (stable states) are studied. A recent work [27] deals with singular perturbations techniques for travelling waves connecting an unstable and a stable state; the stability of these waves for very small γ is also established. Furthermore, in the context of singular perturbation theory, travelling waves for higher order parabolic equations have been studied in [15].

The objective of this paper is to obtain existence results for a large range of parameter values. We therefore study travelling waves of (1.1) via topological arguments rather than perturbation methods. To illustrate the underlying ideas of the method, let us consider the related second order parabolic equation, i.e., $\gamma = 0$. Such equations arise as models in, for example, population genetics and combustion theory [4]. In the special case where $f(u) = f_a(u)$, (1.1) with $\gamma = 0$ admits a travelling wave solution $u(x,t) = \tanh\big(\frac{x+a\sqrt{2}t}{\sqrt{2}}\big)$. This travelling wave connects the two stable homogeneous states $u = -1$ and $u = +1$. The literature on this problem is extensive and we will not attempt to give a complete list. However, a few key references are of importance for explaining the similarities of the second and fourth order problems. In the case $\gamma = 0$ the equation for travelling waves $u(x,t) = U(x + ct)$ is given by $cU' = U'' + f(U)$. A phase plane analysis for both $0 < c \ll 1$ and $c \gg 1$ shows two topologically different phase portraits, from which the conclusion may be drawn that a global bifurcation has to take place for some intermediate c-value(s). In this way a wave speed c_0 can be found for which a travelling wave exists which connects the two local maxima of F. In this context we mention the work by Fife and McLeod [13] based on an analytic approach and Conley's more topological approach [8].

From the second order problem we learn that for the present problem it is sensible to look for topologically different phase portraits (in $\mathbb{R}^4$) for small and large values of c. A big part of our analysis will be to do just that.

In order to simplify the exposition of the main results we reformulate (1.1) as

$$u_t = -u_{xxxx} + \alpha u_{xx} + f(u), \tag{1.2}$$

via the rescaling $x \mapsto \gamma^{\frac{1}{4}} x$, with $\alpha = \frac{1}{\sqrt{\gamma}}$. Notice that (1.2) also has meaning for $\alpha \leq 0$.

Let us start now with the hypotheses on the nonlinearity:

$$(\mathrm{H}_0)\quad \begin{cases} \bullet\ F'(u) = f(u) \in C^1(\mathbb{R}); \\ \bullet\ f(u) = 0 \Leftrightarrow u \in \{\pm 1, -a\} \text{ for some } a \in (-1,1), \text{ and } f'(\pm 1) \neq 0,\ f'(-a) \neq 0; \\ \bullet\ F(-1) < F(+1); \\ \bullet\ F(u) \to -\infty \text{ as } u \to \pm\infty; \\ \bullet\ \text{for some } M > 0 \text{ it holds that } f'(u) \leq M \text{ for all } u \in \mathbb{R}.^2 \end{cases}$$

Of course, the prototypical example $f_a(u) = (u+a)(1-u^2)$ satisfies (H_0). We remark that the third condition excludes the existence of a standing wave which connects

[2]Note that $f'(u)$ may be unbounded from below.

two different equilibria. The last condition is a technical one, which we use to obtain certain a priori bounds. Without loss of generality we set

$$F(u) = \int_1^u f(s)ds,$$

so that $F(1) = 0$.

Denote the wave speed by c, and, searching for a travelling wave, we set $u(x,t) = U(x+ct)$, which, switching to lower case again, reduces (1.2) to the ordinary differential equation

$$cu' = -u'''' + \alpha u'' + f(u). \tag{1.3}$$

An important ingredient of our analysis is a conserved quantity for (1.3) when $c = 0$, which is a Lyapunov function when $c \neq 0$. Define

$$\mathcal{E}(u, u', u'', u''') \stackrel{\text{def}}{=} -u'u''' + \frac{1}{2}u''^2 + \frac{\alpha}{2}u'^2 + F(u). \tag{1.4}$$

Multiplying (1.3) by u' we find that

$$\mathcal{E}'(u, u', u'', u''') = cu'^2, \tag{1.5}$$

so that $\mathcal{E}$, which will be referred to as the *energy* of the solution, is increasing along orbits if $c > 0$, constant if $c = 0$, and decreasing if $c < 0$. When we are looking for a solution of (1.3) connecting $u = -1$ to $u = 1$, we see that we can restrict our attention to $c > 0$.

The first theorem deals with the connection between the two stable states $u = -1$ and $u = +1$. This connection is nongeneric with respect to the wave speed c. Noting that $F(u) \leq 0$ for all $u \in \mathbb{R}$ if f satisfies hypothesis (H_0), we define

$$\sigma(f) \stackrel{\text{def}}{=} \min_{-1<u<-a} \frac{-F(u)}{2f(u)^2}. \tag{1.6}$$

THEOREM 1.1. *Let f satisfy hypothesis (H_0) and let $\alpha > \frac{1}{\sqrt{\sigma(f)}}$. Then, for some wave speed $c = c_0(f) > 0$, there exists a travelling wave solution of (1.2) connecting $u = -1$ to $u = +1$.*

The analogous condition on γ for (1.1) reads $0 < \gamma < \sigma(f)$.

At the minimum in (1.6) the equality $\frac{-F(u)}{2f(u)^2} = \frac{-1}{4f'(u)}$ holds. We easily derive that for our model nonlinearity f_a we have $\sigma(f_a) > \frac{1}{8(1-a)}$ for all $0 < a < 1$. Although this estimate is sharp for $a \to 0$, it is not sharp at all for larger values of a.

For general nonlinearities $f(u)$ satisfying (H_0), a lower bound on σ is

$$\sigma \geq \min\Big\{ \frac{-1}{4f'(u)} \;\Big|\; u \in (-1,-a) \text{ and } f'(u) < 0 \Big\}. \tag{1.7}$$

This estimate is often easier to compute than σ itself, but it is in general a rather blunt estimate. Finally, we remark that the critical value σ is also encountered in the study of homoclinic orbits for $c = 0$ (see [22, Theorem B]). This originates from the similarity of that problem with the proof of Lemma 5.1, which is in fact the only instance in our analysis where γ is required to be smaller than σ.

We do not obtain much insight in the shape of the travelling wave from Theorem 1.1. Because Theorem 1.1 does not give information about the wave speed, it is not known whether the connected equilibrium points are approached monotonically or in an oscillatory manner. The linearized equation around the equilibrium points leads to the following characteristic equation for the eigenvalues: $c\lambda = -\lambda^4 + \alpha\lambda + f'(\pm 1)$. A few conclusions can be drawn from analyzing this equation. It follows that for $\alpha \geq \sqrt{-4f'(1)}$ the travelling wave tends to $+1$ monotonically as $x \to \infty$. Besides, for $\alpha \leq \sqrt{-4f'(-1)}$ the travelling wave tends to -1 in an oscillatory way as $x \to -\infty$. For other cases the behavior in the limits depends on the value of c.

The travelling wave solution found in Theorem 1.1 connects the two maxima of F. Theorem 1.1 can be extended to potentials F having many local extrema, i.e., $f(u)$ having many zeros. In that case we find a travelling wave connecting the global maximum and the second largest local maximum of F. The other conditions on F remain the same, but we also need that $f(u)u < 0$ for large values of $|u|$. The definition of σ in this case is, setting $\max_{u\in\mathbb{R}} F(u) = 0$,

$$\sigma(f) \stackrel{\text{def}}{=} \inf\left\{ \frac{-F(u)}{2f(u)^2} \;\middle|\; u \in \mathbb{R} \text{ and } f(u)f'(u) > 0 \right\}.$$

The travelling wave solution found in Theorem 1.1 connects the two stable states. The following theorems deal with travelling waves connecting the unstable state $u = -a$ to one of the stable states $u = \pm 1$. These theorems also apply to the parameter regime where $\alpha \geq 0$, but for these parameter values we need an additional condition on f:

(H_1) f satisfies (H_0) and $\lim_{|u|\to\infty} \dfrac{f(u)}{u} = -\infty$.

THEOREM 1.2. *Let $\alpha \in \mathbb{R}$ and let f satisfy hypothesis* (H_0) *if $\alpha < 0$ and* (H_1) *if $\alpha \geq 0$.*[3] *Then for every $c > 0$ there exists a travelling wave solution of* (1.2) *connecting $u = -a$ to $u = -1$.*

The limiting behavior of the travelling waves can be determined from the characteristic equations. For $\alpha \geq \sqrt{-4f'(-1)}$ the solution tends to -1 monotonically for $x \to \infty$ regardless of the speed c. On the other hand, for $\alpha < \sqrt{-4f'(-1)}$ the limit behavior is oscillatory for small c and monotonic for large c. The limit behavior near $u = -a$ as $x \to -\infty$ is more complicated. For small c the behavior is generically oscillatory, while for large c the solutions generically tends to $-a$ monotonically. We do not know whether the behavior is indeed generic. However, for $\alpha > \sqrt{12f'(-a)}$ there is an intermediate range of c-values for which the travelling wave certainly tends to $-a$ monotonically.

For general potentials F this result applies to any pair of consecutive nondegenerate extrema u_- (a minimum) and u_+ (a maximum), for which the interval $\big(F(u_-), F(u_+)\big)$ contains no critical values and either u_- or u_+ is the only critical point at level $F(u_\pm)$. The other conditions on F remain the same. The method of proof of Theorem 1.2 requires only one of the two extrema -1 or $-a$ to be nondegenerate.

The next theorem deals with the case of travelling waves from $-a$ to $+1$.

THEOREM 1.3. *Let $\alpha \in \mathbb{R}$ and let f satisfy hypothesis* (H_0) *if $\alpha < 0$ and* (H_1) *if $\alpha \geq 0$. Then there exists a constant $c^*(f) > 0$, such that for every $c > c^*$ there exists a travelling wave solution of* (1.2) *connecting $u = -a$ to $u = +1$.*

[3] The result also holds when $F(-1) = F(+1)$.

Theorem 1.3 extends to general potentials, giving travelling waves between any pair of consecutive nondegenerate extrema u_- (a minimum) and u_+ (a maximum), provided the local minimum $\tilde{u}_-$ on the other side of u_+, if it exists, satisfies $F(\tilde{u}_-) > F(u_-)$. Of course, if the opposite inequality holds then one can exchange u_- and $\tilde{u}_-$. If equality holds, i.e., $F(\tilde{u}_-) = F(u_-)$, then one obtains for every $c > c^*$ a travelling wave connecting either u_- or $\tilde{u}_-$ to u_+. Again, the other conditions on F remain the same.

In certain cases one obtains information about the constant c^* in Theorem 1.3. In that case the situation is very much analogous to the second order equation.

COROLLARY 1.4. *Let f satisfy hypothesis* (H_0) *and let $\alpha > \frac{1}{\sqrt{\sigma(f)}}$. Then there exists a $c^*(f) > 0$, such that c^* is the largest speed for which there exists a travelling wave solution of* (1.2) *connecting $u = -1$ to $u = +1$. Moreover, for all $c > c^*$ there exists a travelling wave solution of* (1.2) *connecting $u = -a$ to $u = +1$.*

Finally, we discuss nonlinearities with different behavior for $u \to \pm\infty$. Assume that f has *two* zeros and satisfies

$$(\mathrm{H}_2)\quad \begin{cases} \bullet\ F'(u) = f(u) \in C^1(\mathbb{R}); \\ \bullet\ f(u) = 0 \Leftrightarrow u \in \{0, 1\}, \text{ and } f'(0) \neq 0,\ f'(1) \neq 0; \\ \bullet\ \text{for some } D < 0 \text{ it holds that } F(u) > F(1) \text{ for all } u < D; \\ \bullet\ F(u) \to -\infty \text{ as } u \to \infty; \\ \bullet\ \text{if } \alpha \geq 0, \text{ then } \lim_{|u|\to\infty} \frac{f(u)}{u} = -\infty. \end{cases}$$

A typical example is $f(u) = u(1 - u)$. The following theorem is analogous to Theorem 1.2.

THEOREM 1.5. *Let $\alpha \in \mathbb{R}$ and let f satisfy hypothesis* (H_2). *Then for every $c > 0$ there exists a travelling wave solution of* (1.2) *connecting $u = 0$ to $u = 1$.*

This last theorem is just an example of how the methods in this paper can also be applied when $F(u)$ does not tend to $-\infty$ as $u \to \pm\infty$. The theorem holds under weaker conditions, but we leave this to the interested reader.

Of the results in this paper, the proof of Theorem 1.1 is by far the most involved. This is caused by the fact that connections between local maxima are nongeneric with respect to the wave speed c. Hence, part of the problem is to determine the wave speed c. The idea behind the proof is that one can detect a change in the phase portrait (in $\mathbb{R}^4$) of (1.3) as c goes from small values to large values. In particular, looking for a travelling wave which connects -1 to $+1$, we investigate the global behavior of the orbits in the stable manifold $W^s(1)$ of the equilibrium point $u = +1$.

The analysis for $c > 0$ large is based on a continuation argument deforming the nonlinearity $f(u)$ into a function which is linear on some interval containing $u = 1$.

For $c > 0$ small the analysis is much more involved. A crucial step is that for $c = 0$ all orbits in $W^s(1)$ are unbounded. A first result in this direction was already proved in [29]. There it was shown that, for γ not too large, the bounded stationary solutions of (1.1) correspond exactly to the bounded stationary solutions of the second order equation ($\gamma = 0$). This excludes the existence of bounded orbits in $W^s(1)$. However, since the analysis comprises *all* bounded solutions, this result is limited to a restricted parameter regime. In particular, the equilibrium points $u = \pm 1$ need to be real saddles. In the present situation we want to exclude bounded solutions in the stable manifold of $u = 1$, i.e., we can restrict the analysis to the energy level $\mathcal{E} = 0$. This allows us to cover a larger range of α-values; to be precise, $\alpha > \frac{1}{\sqrt{\sigma(f)}}$. This parameter regime includes cases where both equilibrium points $u = \pm 1$ are saddle-foci. To give an example, for our model nonlinearity $f_a = (u + a)(1 - u^2)$ with $0 < a < 1$

the result from [29] holds for $\alpha \geq \sqrt{8(1+a)}$. The equilibrium points $u = 1$ and $u = -1$ become saddle-foci for $\alpha < \sqrt{8(1+a)}$ and $\alpha < \sqrt{8(1-a)}$, respectively. One may compare this to the estimate $\sigma(f_a) > \frac{1}{8(1-a)}$. Notice that this estimate, although sharp for $a \to 0$, is very blunt for a close to 1.

For the description of unbounded orbits we use a modified Poincaré transformation which we believe is of independent interest. We investigate the unbounded orbits, and we will show that, in an appropriate compactification of the phase space, these orbits must converge to a unique periodic orbit lying at infinity in the phase space. The analysis at infinity largely relies on a global analysis of bounded and unbounded solutions of the family of equations

$$u'''' + u^s = 0 \quad \text{with the convention that} \quad u^s = |u|^{s-1}u, \ s \geq 1.$$

This equation is invariant under the scaling $u(t) \mapsto \kappa u(\kappa^{\frac{s-1}{4}}t)$ for all $\kappa > 0$. The analysis of this equation is in particular used in the proof of finite time blow-up of unbounded solutions, and, more importantly, to determine the behavior of unbounded orbits for $0 \leq c \ll 1$.

From this analysis we conclude that the phase portrait for c positive but small is different from the phase portrait for c large, which in turn is used to prove the existence of a connection between -1 and $+1$ for some intermediate wave speed c_0.

The organization of the paper is as follows. We start with some a priori bounds in section 2. In section 3 we give the proof of Theorem 1.1, and in sections 4 to 6 the details of this proof are filled in. In particular, in section 4 we perform an analysis of the flow "at infinity." Sections 5 and 6 deal with the analysis of the orbits in $W^s(1)$ for small c and large c, respectively. Section 7 discusses the existence of travelling waves connecting $u = -a$ to $u = \pm 1$; Theorems 1.2 to 1.5 are proved here. We conclude with some remarks on open problems in section 8.

2. A priori estimates. We establish a priori bounds on the wave speed c and the profile u for any travelling wave connecting -1 and $+1$. The bound on the wave speed c holds for all $\alpha \in \mathbb{R}$.

LEMMA 2.1. *Let f satisfy hypothesis* (H_0) *and let $\alpha \in \mathbb{R}$. There exists a constant c_0, depending only on α, $F(-1)$, $F(-a)$, and the upper bound M for $f'(u)$, such that when $c > 0$ is a speed for which there exists a travelling wave solution of* (1.3) *connecting -1 to $+1$, then $c \leq c_0$.*

Proof. Suppose u is a solution of (1.3) connecting -1 to $+1$. Integrating (1.5), we have

$$-F(-1) = F(1) - F(-1) = c\int_{-\infty}^{\infty} u'^2. \tag{2.1}$$

Multiplying (1.3) by u'' and integrating (by parts) we obtain

$$\int_{-\infty}^{\infty} u'''^2 + \alpha\int_{-\infty}^{\infty} u''^2 = \int_{-\infty}^{\infty} (f(u))'u' = \int_{-\infty}^{\infty} f'(u)u'^2 \leq M\int_{-\infty}^{\infty} u'^2 = M\frac{-F(-1)}{c}. \tag{2.2}$$

Let $u_1 \in (-a, 1)$ be defined by

$$F(u_1) = \frac{F(-a) + F(-1)}{2}.$$

There must be points $t_0, t_1 \in \mathbb{R}$, $t_0 < t_1$, such that $u(t_0) = -a$, $u(t_1) = u_1$, and $u(t) \in [-a, u_1]$ for $t \in [t_0, t_1]$. The length of this interval is estimated from below by

$$(u_1 + a)^2 = \left(\int_{t_0}^{t_1} u'(t) dt \right)^2 \leq (t_1 - t_0)^2 \int_{t_0}^{t_1} u'(t)^2 dt \leq (t_1 - t_0)^2 \frac{-F(-1)}{c}.$$

On the one hand, because the energy $\mathcal{E}$ increases along orbits, we have

$$\begin{aligned} \int_{t_0}^{t_1} \left(-u'''(t)u'(t) + \frac{1}{2}u''(t)^2 + \frac{\alpha}{2}u'(t)^2 \right) dt & \\ \geq \int_{t_0}^{t_1} \big(F(-1) - F(u(t)) \big) dt & \\ \geq (F(-1) - F(u_1))(t_1 - t_0) = \frac{F(-1) - F(-a)}{2}(t_1 - t_0) & \\ \geq \frac{F(-1) - F(-a)}{2}(u_1 + a)\sqrt{\frac{c}{-F(-1)}}. & \end{aligned} \tag{2.3}$$

We now first restrict to the case that $\alpha > 0$, and come back to the other case later on. Using (2.1) and (2.2), we obtain the estimate

$$\begin{aligned} \int_{t_0}^{t_1} \left(-u'''(t)u'(t) + \frac{1}{2}u''(t)^2 + \frac{\alpha}{2}u'(t)^2 \right) dt & \\ \leq \int_{t_0}^{t_1} \left(\frac{1}{2}\big(u'''(t)^2 + u''(t)^2 \big) + \frac{1+\alpha}{2}u'(t)^2 \right) dt & \\ \leq \left(M \max\left\{ \frac{1}{\alpha}, 1 \right\} + 1 + \alpha \right) \frac{-F(-1)}{2c}. & \end{aligned} \tag{2.4}$$

By combining (2.3) and (2.4) we obtain

$$\frac{F(-1) - F(-a)}{2}(u_1 + a)\sqrt{\frac{c}{-F(-1)}} \leq \left(M \max\left\{ \frac{1}{\alpha}, 1 \right\} + 1 + \alpha \right) \frac{-F(-1)}{2c}.$$

Since also

$$\frac{F(-1) - F(-a)}{2} = F(u_1) - F(-a) \leq \frac{M}{2}(u_1 + a)^2,$$

it follows that

$$c \leq M^{\frac{1}{3}} \left(M \max\left\{ \frac{1}{\alpha}, 1 \right\} + 1 + \alpha \right)^{\frac{2}{3}} \frac{-F(-1)}{F(-1) - F(-a)}.$$

This completes the proof of the lemma for the case that $\alpha > 0$.

We now deal with the case $\alpha \leq 0$. The first part of estimate 2.4 is replaced by

$$\begin{aligned} \int_{t_0}^{t_1} \left(-u'''(t)u'(t) + \frac{1}{2}u''(t)^2 + \frac{\alpha}{2}u'(t)^2 \right) dt & \\ \leq \int_{-\infty}^{\infty} \left(\frac{1}{2}u'''(t)^2 + \frac{1}{2}u''(t)^2 + \frac{1}{2}u'(t)^2 \right) dt & \end{aligned}$$

$$= \int_{-\infty}^{\infty} \left(u'''(t)^2 + \alpha u''(t)^2 + \left(\frac{1}{2} - \alpha\right) u''(t)^2 - \frac{1}{2} u'''(t)^2 + \frac{1}{2} u'(t)^2 \right) dt$$
$$\leq \int_{-\infty}^{\infty} \left(u'''(t)^2 + \alpha u''(t)^2 + \frac{4\alpha^2 - 4\alpha + 5}{8} u'(t)^2 \right) dt,$$

where we have used that $\int_{-\infty}^{\infty} u''^2 \leq \lambda \int_{-\infty}^{\infty} u'''^2 + \frac{1}{4\lambda} \int_{-\infty}^{\infty} u'^2$ for all $\lambda > 0$. The remainder of the proof is the same as above. □

The L^∞-bound on the profile u holds for $\alpha > 0$, or equivalently, for all $\gamma > 0$.

LEMMA 2.2. *Let f satisfy hypothesis* (H_0) *and let $\alpha > 0$. There exists a constant C_1, depending only on α, $F(-1)$, $F(-a)$, and the upper bound M for $f'(u)$, such that when u is, for some $c > 0$, a travelling wave solution of* (1.3) *connecting -1 to $+1$, then $F(u) \geq C_1$.*

Proof. We may suppose that there is a connection u with range not contained in the bounded interval $\{u \in \mathbb{R} \,|\, F(u) \geq F(-a)\}$, otherwise we already have our desired uniform bound. Therefore, without loss of generality we may assume that

$$F(u(0)) = \min_{t \in \mathbb{R}} F(u(t)) < F(-a). \tag{2.5}$$

We consider the case where $u(0) < -1$ (the case $u(0) > 1$ is completely analogous). Since

$$\mathcal{E}(u, u', u'', u''')(t) \in \big(F(-1), F(1)\big) = \big(F(-1), 0\big) \quad \text{for all } t \in \mathbb{R}, \tag{2.6}$$

we clearly have that

$$u(0) < -1, \quad u'(0) = 0, \quad 0 < \sqrt{2\big(F(-1) - F(u(0))\big)} < u''(0) < \sqrt{-2F(u(0))}.$$

We now consider two cases: $u'''(0) \geq 0$ and $u'''(0) < 0$. We start with the latter case. Since $u(t)$ tends to an equilibrium point as $t \to -\infty$, there exists a $t_1 < 0$ such that $u'''(t) < 0$ for $t_1 < t < 0$ and $u'''(t_1) = 0$, (1.5) implies that

$$-u'(t)u'''(t) + F(u(t)) - F(u(0)) = -\frac{1}{2}\big(u''(t)^2 - u''(0)^2\big) - \frac{\alpha}{2} u'(t)^2 + c \int_0^t u'(s)^2 ds. \tag{2.7}$$

By (2.5) we know that $F(u(t_1)) \geq F(u(0))$, so that

$$\frac{1}{2}\big(u''(t_1)^2 - u''(0)^2\big) + \frac{\alpha}{2} u'(t_1)^2 \leq -c \int_{t_1}^0 u'(s)^2 ds.$$

Since $u''(t)$ decreases on $(t_1, 0)$ and α is positive, this implies that $c < 0$, a contradiction.

We now deal with the case that $u'''(0) \geq 0$. Since $u''''(0) > 0$ by the differential equation, and since $u(t)$ tends to an equilibrium point as $t \to \infty$, there exists a $t_2 > 0$ such that $u'''(t) > 0$ for $0 < t < t_2$ and $u'''(t_2) = 0$. By (2.5) we know that $F(u(t_2)) \geq F(u(0))$. Since $\alpha > 0$, it follows from (2.7) that

$$\frac{\alpha}{2} u'(t_2)^2 \leq c \int_0^{t_2} u'(s)^2 ds \leq c \int_{-\infty}^{\infty} u'(s)^2 ds \leq -F(-1). \tag{2.8}$$

Furthermore, from the fact that $u''(t)$ increases on $(0, t_2)$ we infer that

$$u''(0)t \leq u'(t) \leq u'(t_2) \qquad \text{for } t \in [0, t_2]. \tag{2.9}$$

On the one hand it follows from (2.8) and (2.9) that $\frac{\alpha}{2}u'(t_2)^2 \le c\int_0^{t_2} u'(s)^2 ds \le cu'(t_2)^2 t_2$, hence

$$t_2 \ge \frac{\alpha}{2c}. \tag{2.10}$$

On the other hand it follows from (2.8) and (2.9) that $-F(-1) \ge c\int_0^{t_2} u'(s)^2 ds \ge \frac{1}{3}ct_2^3 u''(0)^2$. Combining with (2.10) we thus obtain that

$$u''(0)^2 \le \frac{-24c^2F(-1)}{\alpha^3}.$$

This gives a bound on $u''(0)^2$, because it follows from Lemma 2.1 that the wave speed c is bounded above by a constant $c_0\big(\alpha, M, F(-a), F(-1)\big)$.

Finally, by (2.5) and (2.6) we have

$$F(u(t)) \ge F(u(0)) \ge F(-1) - \frac{1}{2}u''(0)^2 \qquad \text{for all } t \in \mathbb{R}.$$

This completes the proof of Lemma 2.2. □

3. Proof of Theorem 1.1. In this section we give the proof of Theorem 1.1. Some of the major steps, which require a quite involved analysis, are only stated as a proposition in this section and are proved in subsequent sections.

We first use the a priori bounds of section 2 to reduce our analysis to nonlinearities $f(u)$ of the form $f(u) = -u^3 + g(u)$, where $g(u)$ has compact support. The advantage of such nonlinearities is that they behave nicely as $u \to \pm\infty$, and it will thus be possible to analyze the flow near/at infinity.

Let $f(u)$ satisfy hypothesis (H_0). Lemma 2.2 implies that there exists a constant C_0 such that any travelling wave solution u connecting -1 to $+1$ satisfies $\|u\|_\infty < C_0$. Define the cut-off function $\phi \in C_0^\infty$ with $0 \le \phi \le 1$, $\phi(y) = 1$ for $|y| \le C_0$, and $\phi(y) = 0$ for $|y| > C_0 + 1$. We now consider the modified nonlinearity $\tilde{f}(u) = \phi(u)f(u) - u^3(1-\phi(u))$. Lemma 2.2 ensures that u is a travelling wave solution for nonlinearity $f(u)$ if and only if u is a travelling wave solution for nonlinearity $\tilde{f}(u)$. Besides, $\sigma(f) = \sigma(\tilde{f})$. This shows that we may restrict our analysis to nonlinearities $f(u)$ such that

(3.1) $f(u) = -u^3 + g(u)$ *with* g *compactly supported, and* f *satisfies hypothesis* (H_0).

The purpose of the reduction to nonlinearities f which satisfy (3.1) is that it makes it possible to analyze the orbits which are unbounded. An important property of unbounded solutions, which we will need in the following, is formulated in the next lemma.

LEMMA 3.1. *Let* f *satisfy hypothesis* (3.1) *and let* $\alpha, c \in \mathbb{R}$. *Then any unbounded solution of* (1.3) *blows up in finite time.*

This lemma is proved in section 4.5, Theorem 4.8(b), and is based on the analysis of the flow near/at infinity.

As already discussed in the introduction, denote the wave speed by c. For finding a travelling wave we set $u(x,t) = U(x+ct)$, which reduces (1.1) to the ordinary differential equation (1.3). Written as a four-dimensional system, (1.3) becomes

$$u' = v; \quad v' = w; \quad w' = z; \quad z' = \alpha w - cv + f(u). \tag{3.2}$$

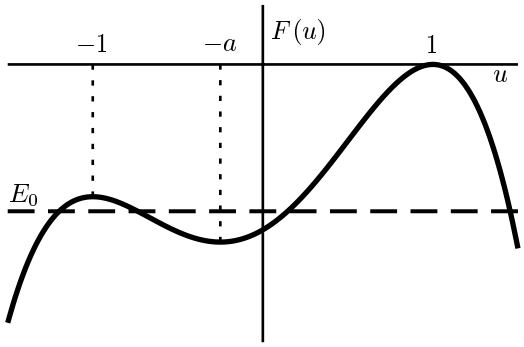


FIG. 3.1. *The potential $F(u)$ and the energy level E_0 separating $u = -a$ from $u = \pm 1$.*

The equilibria of this system are $(u, v, w, z) = (-1, 0, 0, 0)$, $(u, v, w, z) = (-a, 0, 0, 0)$, and $(u, v, w, z) = (1, 0, 0, 0)$ (for short: $u = -1$, $u = -a$, and $u = 1$). To prove Theorem 1.1 we look for a $c \neq 0$ and a corresponding heteroclinic orbit of (3.2) connecting $u = -1$ to $u = 1$. Linearizing around $u = \pm 1$ we find that, irrespective of c, both $u = -1$ and $u = 1$ have two-dimensional stable and unstable manifolds, denoted by $W^s(\pm 1)$ and $W^u(\pm 1)$. Generically $W^s(1)$ and $W^u(-1)$ will not intersect but varying c we expect to pick up a nonempty intersection.

We recall that the *energy* is defined as

$$\mathcal{E}(u, v, w, z) \stackrel{\text{def}}{=} -vz + \frac{1}{2}w^2 + \frac{\alpha}{2}v^2 + F(u),$$

where the potential $F(u) = \int_1^u f(s)ds$ is depicted in Figure 3.1. Since we are looking for a solution of (1.3) which connects $u = -1$ to $u = 1$, we see from (1.5) that we can restrict our attention to $c > 0$. The energy $\mathcal{E}$ thus increases along orbits.

To separate the equilibrium point $u = -a$ from $u = \pm 1$, we choose an energy level E_0 such that (see also Figure 3.1)

$$F(-a) < E_0 < F(-1) < 0,$$

and we define the set

$$K \stackrel{\text{def}}{=} \{(u, v, w, z) \in \mathbb{R}^4 \,|\, \mathcal{E}(u, v, w, z) \geq E_0\}. \tag{3.3}$$

This allows us to formulate the following lemma.

LEMMA 3.2. *Let f satisfy hypothesis* (3.1) *and let $\alpha \in \mathbb{R}$. If $c > 0$ is such that $W^s(1) \cap W^u(-1) = \emptyset$, then every orbit in $W^s(1)$ enters K through its boundary δK and $\hat{\Gamma} = W^s(1) \cap \delta K$ is a simple closed curve. The set of positive c for which this property holds is open and $\hat{\Gamma}$ varies continuously with c.*

Proof. In view of (1.5) the intersection of $W^s(1)$ and δK must be transversal. Assume that $W^s(1) \cap W^u(-1) = \emptyset$. We need to show that every orbit in $W^s(1)$ can be traced back to δK, for then there is bijection between $W^s(1) \cap \delta K$ and a smooth simple closed curve in $W^s_{\text{loc}}(1)$ winding around $u = 1$ (in $W^s_{\text{loc}}(1)$). Arguing by contradiction we assume that there is an orbit in $W^s(1)$ which is completely contained in K. Let $u(t)$ be a solution representing this orbit. Then $u(t)$ exists on some maximal time interval $(t_{\min}, \infty)$. Since $u(t)$ has energy larger than E_0, it follows from (1.5) and

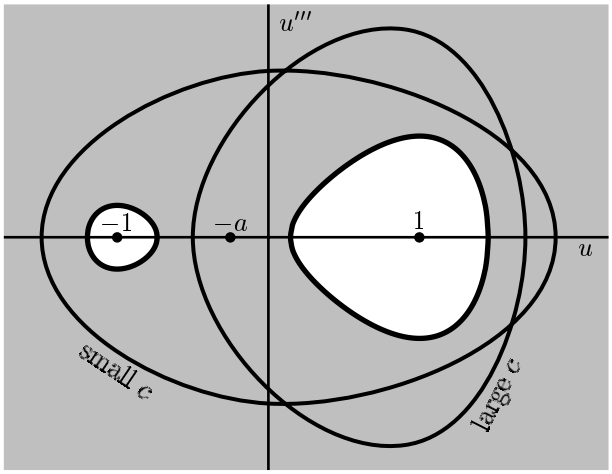


FIG. 3.2. *The projection (in grey) of δK onto the (u,z)-plane. The closed curves which form the boundary of the grey area are given by* (3.6). *The other two curves depict Γ (i.e., the projection of $W^s(1)\cap\delta K$ onto the (u,z)-plane) for small c and large c.*

(3.3) that

$$\int_{t_{\min}}^{\infty} u'^2 \le \frac{F(1)-E_0}{c} = \frac{-E_0}{c}, \tag{3.4}$$

so that $u(t)$ remains bounded on $(t_{\min},\infty)$ if $t_{\min}$ is finite. Thus $t_{\min}=-\infty$ and, by Lemma 3.1, $u(t)$ is bounded. It follows from standard arguments that the orbit converges to a limit as $t\to-\infty$. Because $u=-1$ is the only equilibrium in K with energy less than the energy of $u=1$, we infer that $u(t)\in W^u(-1)$. This contradicts the assumption that $W^s(1)\cap W^u(-1)=\emptyset$. The second statement is an immediate consequence of the (topological) transversality of $W^s(1)\cap\delta K$. □

It now suffices to show that there is a $c>0$ for which the assumption of Lemma 3.2 fails. Again arguing by contradiction, we assume that Lemma 3.2 applies to all $c>0$ and search for a topological obstruction. This requires a description of δK that allows us to form a global picture of this set. To this end we write δK as (with $\alpha>0$)

$$\delta K = \left\{ (u,v,w,z)\in\mathbb{R}^4 \;\middle|\; \frac{\alpha}{2}\left(v-\frac{1}{\alpha}z\right)^2+\frac{1}{2}w^2 = E_0-F(u)+\frac{1}{2\alpha}z^2 \right\}. \tag{3.5}$$

In Figure 3.2 we have plotted the projection of δK onto the (u,z)-plane. For (u,z) lying inside one of the two closed curves (see Figure 3.2) defined by

$$E_0-F(u)+\frac{1}{2\alpha}z^2=0, \tag{3.6}$$

every (u,v,w,z) belongs to K, hence there are no points in δK with (u,z) lying inside these two closed curves. For (u,z) lying outside the two closed curves we have that (u,v,w,z) is in K if (v,w) is outside the ellipse defined by $\frac{\alpha}{2}\big(v-\frac{1}{\alpha}z\big)^2+\frac{1}{2}w^2=0$. We conclude that the projection of δK onto the (u,z)-plane is the region outside the two closed curves defined by (3.6); see Figure 3.2.

The projection of δK onto the (u,z)-plane maps $\hat{\Gamma} = W^s(1) \cap \delta K$, which by assumption exists for all $c > 0$, to a closed but not necessarily simple curve Γ in the (u,z)-plane for which the winding numbers[4] $n(\Gamma,-1)$ and $n(\Gamma,1)$ around $(u,z) = (-1,0)$ and $(u,z) = (1,0)$, respectively, are well defined and independent of c (by continuity). However, the following proposition establishes the configuration depicted in Figure 3.2, contradicting the assumption that $W^s(1) \cap W^u(-1) = \emptyset$ for all $c > 0$, and thereby completing the proof of Theorem 1.1.

PROPOSITION 3.3. *Let f satisfy hypothesis* (3.1).

(a) *Let $\alpha > \frac{1}{\sqrt{\sigma(f)}}$. Then there exists a $c_* > 0$ such that $n(\Gamma,-1) = 1$ and $n(\Gamma,1) = 1$ for all $0 < c < c_*$.*

(b) *Let $\alpha \in \mathbb{R}$. Then there exists a $c^* > 0$ such that $n(\Gamma,-1) = 0$ and $n(\Gamma,1) = 1$ for all $c > c^*$.*

Part (a) of Proposition 3.3 will be proved in Theorem 5.3 in section 5, while part (b) is proved in section 6, Theorem 6.1.

4. Classification of unbounded solutions. In this section we investigate the behavior of unbounded solutions, or in other words, we analyze the flow at infinity. This analysis is relevant both for the proof of finite time blow-up of unbounded solutions, and to determine the behavior of unbounded orbits for $0 \le c \ll 1$. We have argued in section 3 that we may restrict our attention to nonlinearities of the form $f(u) = -u^3 + g(u)$, where $g(u)$ has compact support. It turns out that the flow for large u is governed by the *reduced* equation $u'''' + u^3 = 0$, i.e., only the highest order derivative and the highest order term in the nonlinearity play a role at infinity. In the following sections we investigate the reduced equation, and in section 4.5 we come back to the full equation.

4.1. A modified Poincaré transformation. We analyze the reduced equation

$$u'''' + u^s = 0 \quad \text{with the convention that} \quad u^s = |u|^{s-1}u, \ s \ge 1, \tag{4.1}$$

and we use this notational convention throughout. Written as a system, (4.1) reads

$$x_1' = x_2; \quad x_2' = x_3; \quad x_3' = x_4; \quad x_4' = -x_1^s, \tag{4.2}$$

where x_1, x_2, x_3, and x_4 correspond to u, u', u'', and u'''. Note that for this system the energy (or Hamiltonian)

$$H(x_1,x_2,x_3,x_4) \stackrel{\text{def}}{=} -x_2x_4 + \frac{x_3^2}{2} - \frac{|x_1|^{s+1}}{s+1} \tag{4.3}$$

is a conserved quantity.

Introduce five new dependent variables X_1, X_2, X_3, X_4, and $X_5 > 0$ by setting

$$x_i = \frac{X_i}{X_5^{a_i}} \qquad (i = 1,2,3,4), \tag{4.4}$$

where the exponents a_i are to be chosen shortly. Unbounded orbits of (4.2) will correspond to orbits in the new variables with X_5 approaching zero. By substituting

[4] We may choose the orientation of the simple closed curve in $W^s_{\text{loc}}(1)$ winding around $u = 1$ in such a way that its projection onto the (u,z)-plane has winding number equal to $+1$.

(4.4) in (4.2) we obtain the equations

$$X_5X_1' - a_1X_1X_5' = X_2X_5^{1+a_1-a_2}; \tag{4.5a}$$
$$X_5X_2' - a_2X_2X_5' = X_3X_5^{1+a_2-a_3}; \tag{4.5b}$$
$$X_5X_3' - a_3X_3X_5' = X_4X_5^{1+a_3-a_4}; \tag{4.5c}$$
$$X_5X_4' - a_4X_4X_5' = -X_1^sX_5^{1+a_4-sa_1}, \tag{4.5d}$$

with a fifth equation pending. We choose the exponents in such a way that all the exponents in the right-hand sides of (4.5) are the same, i.e.,

$$b \stackrel{\text{def}}{=} 1 + a_1 - a_2 = 1 + a_2 - a_3 = 1 + a_3 - a_4 = 1 + a_4 - sa_1.$$

Solving for a_1, a_2, a_3, a_4, and b we find

$$a_1 = 4\lambda;\ a_2 = (s+3)\lambda;\ a_3 = (2s+2)\lambda;\ a_4 = (3s+1)\lambda;\ b = 1-(s-1)\lambda, \tag{4.6}$$

where λ is still free and, for the moment, positive. We close system (4.5) by imposing as a fifth equation

$$X_1^sX_1' + X_2X_2' + X_3X_3' + X_4X_4' = 0. \tag{4.7}$$

If we multiply (4.5a)–(4.5d) by X_1^s, X_2, X_3, and X_4 respectively, and add up the resulting equations, we obtain

$$PX_5' = -\frac{1}{\lambda}QX_5^b. \tag{4.8}$$

Here we have set

$$P \stackrel{\text{def}}{=} 4|X_1|^{s+1} + (3+s)X_2^2 + (2+2s)X_3^2 + (1+3s)X_4^2, \tag{4.9}$$

which is nonnegative, and

$$Q \stackrel{\text{def}}{=} X_1^s(X_2 - X_4) + X_3(X_2 + X_4).$$

Introducing a new independent variable, we write

$$\dot{X}_5 = PX_5^{(s-1)\lambda}X_5' = -\frac{1}{\lambda}QX_5, \tag{4.10}$$

where the dot denotes derivation with respect to this new independent variable from which the old one may be recovered by integration. Thus, combining (4.10) and (4.5), we arrive at the system

$$\dot{X}_1 = X_2P - 4X_1Q\,; \tag{4.11a}$$
$$\dot{X}_2 = X_3P - (3+s)X_2Q\,; \tag{4.11b}$$
$$\dot{X}_3 = X_4P - (2+2s)X_3Q\,; \tag{4.11c}$$
$$\dot{X}_4 = -X_1^sP - (1+3s)X_4Q\,. \tag{4.11d}$$

Note that X_5 has been decoupled from the equations. By construction the system (4.11) leaves the surfaces

$$\Sigma \stackrel{\text{def}}{=} \left\{(X_1,X_2,X_3,X_4) \;\middle|\; \frac{|X_1|^{s+1}}{s+1} + \frac{X_2^2}{2} + \frac{X_3^2}{2} + \frac{X_4^2}{2} = C_0\right\} \cong S^3 \tag{4.12}$$

invariant for all $C_0 > 0$. The free parameter λ appears only in (4.10) and may be discarded.

The Poincaré transformation (4.4) is used here to blow up the flow near "infinity." As will be explained in section 4.4 this is equivalent to blowing up the flow near the equilibrium point $u = 0$. This blowing-up technique is frequently used in the study of flows in the neighborhood of nonhyperbolic equilibrium points (see, e.g., [10, 11, 21]). The transformation defined by (4.4) and (4.12) is a variant of the standard Poincaré transformation, which has $a_1 = a_2 = a_3 = a_4 = 1$ and imposes as fifth equation that $X_1^2 + X_2^2 + X_3^2 + X_4^2 + X_5^2$ be constant, so that the transformed problem is situated on the Poincaré sphere. The modification presented above, in particular the choice of exponents, is needed to obtain a nontrivial vector field at infinity from which we may derive the qualitative properties of the flow of the system (4.2) near infinity. The values of the exponents are derived from the invariance of (4.1) under the scaling $u(t) \mapsto \kappa u(\kappa^{\frac{s-1}{4}} t)$.

In (4.7) we have chosen not to include a term $X_5 X_5'$ and to modify the exponent of X_1. This simplifies the new vector field and allows the decoupling of the $\dot{X}_5$-equation. Note that instead of a Poincaré sphere we now have a Poincaré cylinder Π, namely, the topological product of the deformed sphere Σ and the positive X_5-axis:

$$\Pi \stackrel{\text{def}}{=} \{(X_1, X_2, X_3, X_4, X_5) \,|\, (X_1, X_2, X_3, X_4) \in \Sigma,\ X_5 \geq 0\} \cong S^3 \times [o, \infty).$$

The flow of (4.2) is completely determined by the flow of (4.11) on Σ. Therefore, we have a reduction from dimension 4 for (4.2) to dimension 3 for (4.11). The role of $X_5 = 0$ and $X_5 = \infty$ can be reversed by changing from positive to negative λ at the expense of a minus sign in (4.10).

Remark 4.1. The choice of $C_0 > 0$ in (4.12) is arbitrary, because the flows on all spheres Σ are C^1-conjugated (modulo the introduction of the new independent variable in (4.10)). This is in fact the very idea of Poincaré transformations, namely, that we divide out the invariance of (4.1) and focus on the resulting flow. From a more abstract point of view one can construct a flow on the quotient manifold $\big(\mathbb{R}^4 \backslash \{0\}\big)/\mathbb{R}^+ \cong S^3$ via the scaling invariance $u(t) \mapsto \kappa u(\kappa^{\frac{s-1}{4}} t)$ ($\mathbb{R}^+$-action); see [20] for more details. Our construction involves explicit choices of coordinates, for which the flows, by general theory, are all related by conjugation.

To be explicit, let X_i and Y_i be two sets of Poincaré coordinates, i.e.,

$$x_i = \frac{X_i}{X_5^{a_i}} = \frac{Y_i}{Y_5^{a_i}} \qquad \text{for } i = 1, 2, 3, 4,$$

with constraints

$$\frac{|X_1|^{s+1}}{s+1} + \frac{X_2^2}{2} + \frac{X_3^2}{2} + \frac{X_4^2}{2} = C_0, \tag{4.13a}$$

$$\frac{|Y_1|^{s+1}}{s+1} + \frac{Y_2^2}{2} + \frac{Y_3^2}{2} + \frac{Y_4^2}{2} = C_1. \tag{4.13b}$$

When we define $\mu = \frac{X_5}{Y_5}$, then the two sets of coordinates are related by

$$X_5 = \mu Y_5 \qquad \text{and} \qquad X_i = \mu^{a_i} Y_i \quad \text{for } i = 1, 2, 3, 4. \tag{4.14}$$

Substituting this into (4.13a) we obtain

$$G(Y_1, Y_2, Y_3, Y_4, \mu) \equiv \mu^{(s+1)a_1} \frac{|Y_1|^{s+1}}{s+1} + \mu^{2a_2} \frac{Y_2^2}{2} + \mu^{2a_3} \frac{Y_3^2}{2} + \mu^{2a_4} \frac{Y_4^2}{2} = C_0.$$

Since $\frac{\partial G}{\partial \mu} > 0$ for all Y_i that obey (4.13b), it follows from the implicit function theorem that $\mu(Y_1, Y_2, Y_3, Y_4)$ is a differentiable function. It is now easily seen from (4.14) that X_i and Y_i are related by a C^1-conjugacy. Therefore, we may choose the constant C_0 according to our liking to obtain a description of the flow that is most suitable to our needs.

4.2. The flow at infinity. For the analysis of (4.11) we first observe the following.

LEMMA 4.2. *System* (4.11) *has no stationary points on* Σ *for any* $C_0 > 0$.

Proof. Since $X_1 = X_2 = X_3 = X_4 = 0$ is excluded we have that P, defined by (4.9), is positive. Equating the right-hand sides of (4.11) to zero and considering the resulting equations as linear equations in P and Q, it follows that we can only have solutions if every determinant of every pair of two equations vanishes. This would give, for instance, that

$$0 \le (2+2s)X_3^2 = (3+s)X_2X_4;$$
$$0 \le 4|X_1|^{s+1} = -(1+3s)X_2X_4.$$

We conclude that $X_2X_4 = 0$ and with any of the $X_i = 0$ the others thus follow immediately. □

We next use the conserved quantity to obtain a further reduction from dimension 3 to dimension 2 for the limit sets of orbits of (4.5) which approach infinity ($X_5 \to 0$) or the origin ($X_5 \to \infty$). In the new variables the Hamiltonian is

$$H = \left(-X_2X_4 + \frac{X_3^2}{2} - \frac{|X_1|^{s+1}}{s+1} \right) X_5^{-4\lambda(s+1)}. \tag{4.15}$$

Denote the first factor of H by H_0:

$$H_0 \stackrel{\text{def}}{=} -X_2X_4 + \frac{X_3^2}{2} - \frac{|X_1|^{s+1}}{s+1}. \tag{4.16}$$

Since H is a conserved quantity, we conclude that for $\lambda > 0$

$$X_5 \to 0 \quad \Leftrightarrow \quad H_0 \to 0. \tag{4.17}$$

For the classification of unbounded orbits we have to analyze the flow restricted to the invariant set given by

$$T \stackrel{\text{def}}{=} \{(X_1, X_2, X_3, X_4) \in \Sigma \mid H_0 = 0\}$$
$$= \left\{ (X_1, X_2, X_3, X_4) \;\middle|\; \frac{|X_1|^{s+1}}{s+1} + \frac{X_2^2}{2} + \frac{X_3^2}{2} + \frac{X_4^2}{2} = C_0, \frac{X_3^2}{2} = X_2X_4 + \frac{|X_1|^{s+1}}{s+1} \right\}.$$

This set is a topological torus as can be seen by setting

$$X_1 = \xi_1; \quad X_2 = \frac{\xi_2 + \xi_4}{\sqrt{2}}; \quad X_3 = \xi_3; \quad X_4 = \frac{\xi_2 - \xi_4}{\sqrt{2}}, \tag{4.18}$$

so that, in terms of the ξ-variables,

$$T = \left\{ (\xi_1, \xi_2, \xi_3, \xi_4) \;\middle|\; \frac{2}{s+1}|\xi_1|^{s+1} + \xi_2^2 = \xi_3^2 + \xi_4^2 = C_0 \right\} \cong S^1 \times S^1. \tag{4.19}$$

Clearly we have that T is the product of two topological circles, one in the (ξ_1, ξ_2)-plane, the other in the (ξ_3, ξ_4)-plane.

LEMMA 4.3. *Let $s \geq 1$ and fix the constant $C_0 > 0$. Then there exist precisely two periodic orbits Λ_- and Λ_+ of (4.11) on the torus T.*

Proof. The proof is based on the observation that the coefficient Q in (4.10), which after transforming by (4.18) reads

$$Q = \sqrt{2}(\xi_1^s \xi_4 + \xi_2 \xi_3), \tag{4.20}$$

plays a double role. Obviously it determines which parts of infinity attract solutions toward $X_5 = 0$, in forward and in backward time. We begin by showing that Q can also be seen as minus the divergence of the vector field restricted to the invariant torus T. From (4.11) and (4.18) we derive

$$\dot{\xi}_1 = \frac{\xi_2 + \xi_4}{\sqrt{2}} P - 4\xi_1 Q\,; \tag{4.21a}$$

$$\dot{\xi}_2 = \frac{\xi_3 - \xi_1^s}{\sqrt{2}} P - ((2+2s)\xi_2 + (1-s)\xi_4) Q\,; \tag{4.21b}$$

$$\dot{\xi}_3 = \frac{\xi_2 - \xi_4}{\sqrt{2}} P - (2+2s)\xi_3 Q\,; \tag{4.21c}$$

$$\dot{\xi}_4 = \frac{\xi_3 + \xi_1^s}{\sqrt{2}} P - ((1-s)\xi_2 + (2+2s)\xi_4) Q\,. \tag{4.21d}$$

We parametrize T by "polar coordinates"

$$\xi_1 = f_1(\phi); \quad \xi_2 = g_1(\phi); \quad \xi_3 = f_2(\theta); \quad \xi_4 = g_2(\theta), \tag{4.22}$$

satisfying

$$f_1' = -g_1; \quad g_1' = f_1^s; \quad f_2' = -g_2; \quad g_2' = f_2. \tag{4.23}$$

Note that when $C_0 = 1$ and $s = 1$ we just have

$$\xi_1 = \cos\phi; \quad \xi_2 = \sin\phi; \quad \xi_3 = \cos\theta; \quad \xi_4 = \sin\theta.$$

From (4.21a), (4.21c), (4.22), and (4.23) we derive that on T the flow is given by

$$\dot{\phi} = \frac{P}{\sqrt{2}} \left(-1 - \frac{g_2}{g_1} \right) + 4Q \frac{f_1}{g_1} \equiv w_1(\phi, \theta); \tag{4.24a}$$

$$\dot{\theta} = \frac{P}{\sqrt{2}} \left(1 - \frac{g_1}{g_2} \right) + 2(s+1)Q \frac{f_2}{g_2} \equiv w_2(\phi, \theta), \tag{4.24b}$$

where in terms of f_1, g_1, f_2, g_2,

$$P = 4(s+1)C_0 + 2(1-s)g_1 g_2, \qquad \text{and} \qquad Q = \sqrt{2}(f_1^s g_2 + f_2 g_1).$$

The functions w_1 and w_2, defined in (4.24), appear to have singularities, but using (4.19) they can be written as

$$w_1(\phi, \theta) = \sqrt{2}\big[-2(s+1)C_0 - (s+3)g_1 g_2 + (s-1)g_2^2 + 4 f_1 f_2\big],$$
$$w_2(\phi, \theta) = \sqrt{2}\big[2(s+1)C_0 - (3s+1)g_1 g_2 + (s-1)g_1^2 + 2(s+1) f_1^s f_2\big].$$

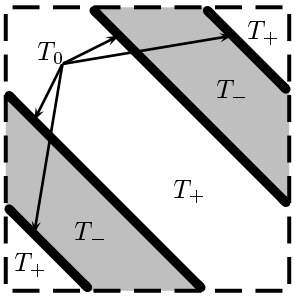


FIG. 4.1. *A fundamental domain of the torus, in which T_-, T_+, and T_0 are indicated (schematically).*

Taking the divergence of the vector field w we obtain (using (4.23),

$$\nabla \cdot w = \frac{\partial w_1}{\partial \phi} + \frac{\partial w_2}{\partial \theta} = \sqrt{2}(-5-3s)(f_1^s g_2 + f_2 g_1) = -(3s+5)Q.$$

Next, we split T into

$$T_+ = \{(X_1, X_2, X_3, X_4) \,|\, Q > 0\} \quad \text{and} \quad T_- = \{(X_1, X_2, X_3, X_4) \,|\, Q < 0\}.$$

These two sets share the boundary

$$T_0 = \{(X_1, X_2, X_3, X_4) \,|\, Q = 0\},$$

which, in view of (4.19) and (4.20), consists of two topological circles, which both wind once around the two homotopically distinct simple loops on the torus (see Figure 4.1). We will show in Lemma 4.4 that, when C_0 is chosen properly, an orbit can only pass through T_0 from T_- to T_+. It then follows from the negativity of $\nabla \cdot w$ in T_+ and the winding properties of T_0 on T that T_+ contains precisely one periodic orbit. The same statement holds for T_- with respect to the backward flow on T.

To be precise, we deduce from (4.22), (4.23), and (4.19) that we may choose $\xi_3 = f_2(\theta) = \sqrt{C_0}\cos\theta$. Define the set $S \stackrel{\text{def}}{=} \{(\theta, \phi) \in T \,|\, \theta = \frac{\pi}{2}\}$, and it follows that

$$\dot{\theta}\Big|_S = \sqrt{2}\left[2(s+1)C_0 - (3s+1)\sqrt{C_0}g_1 + (s-1)g_1^2\right].$$

Since $|g_1| \le \sqrt{C_0}$, it is easy to check that $\dot{\theta}\big|_S \ge 0$, and equality holds only when $g_1 = \sqrt{C_0}$. By continuity arguments the orbit through this point crosses S also in the direction of increasing θ. Thus S is a global section for the flow on T. Moreover, the return map is well defined, since there is no point in T for which the forward orbit is contained in $T \setminus S$. Indeed, such a forward orbit would either be contained in T_- or eventually be in T_+, because T_+ is positively invariant and orbits can only pass through T_0 from T_- to T_+. In the absence of equilibrium points (Lemma 4.2) its ω-limit set would be a periodic orbit. However, there would have to be an equilibrium point inside this periodic orbit, contradicting Lemma 4.2. Hence the return map is well defined. The intersection $S \cap (T_+ \cup T_0)$ consists of the line segment $\{(\theta, \phi) \in T \,|\, \theta = \frac{\pi}{2}, f_1(\phi) \ge 0\}$. The return map maps this line segment into itself, which implies the existence of a periodic orbit in T_+. Similarly, there exists a periodic orbit in T_-. The return map is contracting in T_+ and expanding in T_-, since the divergence

of the vector field is negative in T_+ and positive in T_-. This proves the uniqueness of the two period orbits and shows that all other orbits on the torus T have Λ_- as α-limit set and Λ_+ as ω-limit set.

We remark that the same conclusion can be reached by combining the Poincaré–Bendixson theorem for flows on the torus and the Morse theory for Morse–Smale flows.

Finally, note that, although the preceding proof needs C_0 to have a particular value (see Lemma 4.4 and (4.27)), the statement in Lemma 4.3 is true for any choice of $C_0 > 0$ (see Remark 4.1).

Another observation is that the linear case $s = 1$ may be treated by direct computation, i.e., by transforming the general solution of the then linear equation (4.1) to the X-variables. □

We still have to show that an orbit can only pass through T_0 from T_- to T_+.

LEMMA 4.4. *Let $s > 1$. There exists a $C_0 > 0$ such that orbits on T can only pass through T_0 in the direction from T_- to T_+.*

Proof. We deduce from (4.20) and (4.21) that

$$\dot{Q}\Big|_{Q=0} = P\Big(|\xi_1|^{2s} + \xi_2^2 + \xi_3^2 + \xi_4^2 + (s|\xi_1|^{s-1} - 1)(\xi_2 + \xi_4)\xi_4\Big). \tag{4.25}$$

Notice that for $s = 1$, P is positive on T (see (4.9)), thus $\dot{Q}\big|_{Q=0} > 0$ on T. For $s > 1$ we define R as the second factor in the right-hand side of (4.25) and simplify it using the expression (4.19) for T:

$$\begin{aligned} R &\stackrel{\text{def}}{=} |\xi_1|^{2s} + \xi_2^2 + \xi_3^2 + \xi_4^2 + (s|\xi_1|^{s-1} - 1)(\xi_2 + \xi_4)\xi_4 \\ &= 2C_0 + |\xi_1|^{2s} - \frac{2}{s+1}|\xi_1|^{s+1} - (1 - s|\xi_1|^{s-1})(\xi_2 + \xi_4)\xi_4. \end{aligned} \tag{4.26}$$

From (4.19) we infer that

$$(\xi_2+\xi_4)\xi_4 \le \left(\left(C_0 - \frac{2}{s+1}|\xi_1|^{s+1}\right)^{\frac{1}{2}} + C_0^{\frac{1}{2}}\right)C_0^{\frac{1}{2}} = C_0\left(1+\left(1 - \frac{2}{C_0(s+1)}|\xi_1|^{s+1}\right)^{\frac{1}{2}}\right).$$

Fix

$$C_0 = \frac{2}{s+1}\Big(\frac{1}{s}\Big)^{\frac{s+1}{s-1}}, \tag{4.27}$$

and set

$$|\xi_1| = x\left(\frac{1}{s}\right)^{\frac{1}{s-1}}, \text{ where } 0 \le x \le 1.$$

It follows that

$$\begin{aligned} R &\ge \frac{2}{s+1}\Big(\frac{1}{s}\Big)^{\frac{s+1}{s-1}}\Big(2 + \frac{s+1}{2s}x^{2s} - x^{s+1} - (1 - x^{s-1})(1 + (1 - x^{s+1})^{\frac{1}{2}})\Big) \\ &= \frac{2}{s+1}\Big(\frac{1}{s}\Big)^{\frac{s+1}{s-1}}\Big(1 + \frac{s+1}{2s}x^{2s} - x^{s+1} + x^{s-1} - (1 - x^{s-1})(1 - x^{s+1})^{\frac{1}{2}}\Big) \\ &= \frac{2}{s+1}\Big(\frac{1}{s}\Big)^{\frac{s+1}{s-1}}\Big((1 - x^{s+1})^{\frac{1}{2}}\big((1 - x^{s+1})^{\frac{1}{2}} - (1 - x^{s-1})^{\frac{1}{2}}\big) + x^{s-1} + \frac{s+1}{2}x^{2s}\Big). \end{aligned}$$

Since $0 \le x \le 1$ we see that $R > 0$ unless $x = 0$. Looking at (4.26) we infer that R can only be zero if $\xi_1 = \xi_3 = 0$ and $\xi_2 = \xi_4 = \pm\sqrt{C_0}$, or, in terms of the X_i, if

$X_1 = X_3 = X_4 = 0$. By continuity arguments it follows that also in these two points the orbits go from T_- to T_+. Thus, with the particular choice of C_0 given by (4.27) we have indeed that T_+ is positively invariant and T_- is negatively invariant. □

Having proven the existence of precisely two periodic orbits, Λ_- and Λ_+, on the torus T, we analyze some of their properties.

LEMMA 4.5. *The three nontrivial Floquet multipliers of Λ_+ are contained in the interval $(0,1)$, and the three nontrivial Floquet multipliers of Λ_- are contained in the interval $(1,\infty)$.*

Proof. Restricted to T the nontrivial Floquet multiplier of Λ_+ equals (see, e.g., [26, p. 198])

$$\exp\left(\oint_{\Lambda_+} \nabla \cdot w\right) = \exp\left(\oint_{\Lambda_+} -(3s+5)Q\right).$$

Since Q is uniformly positive on Λ_+, this Floquet multiplier is in $(0,1)$. Close to the periodic orbit Λ_+ we choose ϕ, θ, X_5, and H_0 as coordinates on the Poincaré cylinder Π, where H_0 given by (4.16). Since $H = H_0 X_5^{-4\lambda(s+1)}$ is a conserved quantity on Π, it follows from (4.10) that

$$\dot{H}_0 = -4(s+1)Q\, H_0.$$

Together with (4.10) this implies that the other Floquet multipliers are

$$\exp\left(\oint_{\Lambda_+} -4(s+1)Q\right) \quad \text{and} \quad \exp\left(\oint_{\Lambda_+} -\frac{1}{\lambda}Q\right),$$

which are in $(0,1)$ as before. Thus Λ_+ is exponentially stable. The statement for Λ_- is obtained by time reversal. □

LEMMA 4.6. *Every orbit (other than $\Lambda_\pm$) on the sphere Σ has Λ_- as α-limit set and Λ_+ as ω-limit set.*

Proof. We have already dealt with the flow on the torus T in Lemma 4.3. Orbits of the flow on the complement $\Sigma \backslash T$ of the torus T on the sphere Σ correspond to solutions with nonzero Hamiltonian H. Since X_5 does not appear in (4.10), the motion on Σ is independent of X_5. Let $X_5 \neq 0$, then the dynamics of X_5 are governed by (4.11), and the motion takes place in the part of the Poincaré cylinder Π that corresponds to the finite part of phase space in the x-variables. In other words, orbits of the flow on the set $\Sigma \setminus T$ correspond to solutions of (4.2) with nonzero Hamiltonian.

Since $H = H_0 X_5^{-4\lambda(s+1)}$ and H_0 is bounded on Σ (because Σ is compact), it follows that for such orbits X_5 remains bounded, i.e., in x-variables the solution stays away from the origin. Thus orbits in $\Sigma \setminus T$ are bounded in the X-variables and hence have nonempty invariant α- and ω-limit sets. We have to show that these limit sets can only be the two periodic orbits Λ_- and Λ_+ provided by Lemma 4.3. To this end it suffices to show that all solutions of (4.1) with $H \neq 0$ are unbounded in forward and backward time, i.e., $X_5 \to 0$ along a sequence of points in forward and backward time.

Postponing the proof of the unboundedness of solutions with $H \neq 0$, we first show how unboundedness in backward and forward time implies that Λ_- and Λ_+ are the α- and ω-limit sets. By (4.17) $X_5 \to 0$ implies that also $H_0 \to 0$. An unbounded orbit thus comes arbitrary close to the torus T. We choose an open tubular neighborhood Λ_-^ε of Λ_- in T, such that $Q < 0$ in Λ_-^ε. Clearly all orbits starting in $T \setminus \Lambda_-^\varepsilon$ tend to

Λ_+ in forward time. Note that $T_0 \cup T_+ \subset T \setminus \Lambda_-^\varepsilon$. By compactness of T and since Λ_+ is asymptotically stable (see Lemma 4.5), there exists an open neighborhood T^ε of $T \setminus \Lambda_-^\varepsilon$ in Π such that all orbits starting in T^ε tend to Λ_+ in forward time. Since an orbit which comes close to $X_5 = 0$ (and thus close to T) can only do so with nonnegative Q, it enters T^ε and hence tends to Λ_+. The statement for Λ_- follows by time reversal.

We still have to prove that any solution of (4.1) with nonzero Hamiltonian is unbounded in forward and backward time. We recall that solutions with $H \neq 0$ stay away from the origin. If an orbit would be bounded in backward or forward time, then its (nonempty) α- or ω-limit set would consist of bounded orbits, i.e., orbits which are bounded for all time. However, this is not possible, because it has been proved in [19] that (4.1) admits no bounded solutions except $u \equiv 0$. Here we present a different proof of the fact that (4.1) admits no bounded solutions except $u \equiv 0$, because we need to extend this result to more general situations (see Remark 4.7).

Assume, by contradiction, that $u \not\equiv 0$ is a bounded solution of (4.1). First observe that if u tends to a limit as $t \to \pm\infty$, then this limit can only be 0. It follows that u attains at least one positive maximum or one negative minimum. Switching from u to $-u$ if necessary, we may suppose that u attains a positive maximum at t_0:

$$u(t_0) > 0, \quad u'(t_0) = 0, \quad u''(t_0) \leq 0.$$

Changing from t to $-t$ if necessary, we may assume that $u'''(t_0) \leq 0$ and apply an oscillation argument from [29] which we repeat here for the sake of completeness. There exists a $t^* > t_0$ such that $u'''(t) < 0$ for $t_0 < t < t^*$ and $u'''(t^*) = 0$. Using the fact that

$$H = -u'u''' + \frac{1}{2}u''^2 - \frac{1}{s+1}|u|^{s+1}$$

is constant, it follows that $u(t^*) < -u(t_0)$ and that the next minimum must occur at $t_1 > t^*$ with $u(t_1) < u(t^*) < -u(t_0)$ and both $u''(t_1)$ and $u'''(t_1)$ positive. Repeating this argument we obtain a sequence $t_1 < t_2 < t_3 < \ldots$, in which $u(t)$ has nondegenerate extrema with $|u(t_1)| < |u(t_2)| < |u(t_3)| < \ldots$. By assumption these extrema remain bounded, say $\lim_{i\to\infty} |u(t_i)| = a \in \mathbb{R}^+$, and the derivatives are bounded as well. A compactness argument now shows that there must be a solution $\tilde{u}$ of (4.2) in the ω-limit set of u with

$$\tilde{u}(t_0) = a, \quad \tilde{u}'(t_0) = 0, \quad \tilde{u}''(t_0) < 0, \quad \text{and} \quad \tilde{u}'''(t_0) \leq 0 \qquad \text{at some } t_0 \in \mathbb{R},$$

and such that $|\tilde{u}(t)| \leq a$ for all $t \in \mathbb{R}$. However, when we apply the above argument to $\tilde{u}$ we obtain that $\tilde{u} < -a$ at the first minimum to the right of t_0, a contradiction. This completes the proof of Lemma 4.6. □

Remark 4.7. The oscillation argument above will be applied several times in this paper to differential equations that differ from the present one. It holds that any solution of (1.3) with $c = 0$ and $\beta \geq 0$ which does not have its range contained in

$$\{u \in \mathbb{R} \,|\, F(u) \geq F(-a)\}$$

oscillates toward infinity either in forward or in backward time in exactly the way described above (the additional second order term does not cause any difficulties). For more details we refer to [29].

4.3. The reduced system in the linear limit. We have shown in the previous section that for any $s \geq 1$ the flow of (4.1) is basically governed by two periodic orbits at infinity. For the linear equation ($s = 1$) this was already observed (in a broader setting) by Palis [21]. The analysis thus shows that the behavior for all $s > 1$ is largely analogous to the linear equation. In this section we make some observations about the limit $s \downarrow 1$.

Let us rewrite this system as

$$\dot{X} = V(X; s), \quad X = (X_1, X_2, X_3, X_4). \tag{4.28}$$

Then the vector field $V(\cdot, s)$ is continuously differentiable for every $s \geq 1$ and the first order partial derivatives are bounded on compact sets, uniformly in $s \geq 1$. We do not have that $V(\cdot, s) \to V(\cdot, 1)$ in C^1_{loc} because of the term X_1^s appearing in V, but we do have that $V(\cdot, s) \to V(\cdot, 1)$ uniformly on compact sets. Therefore the orbits of (4.28) with $s > 1$, which are bounded uniformly in s in view of (4.12), converge to orbits of (4.28) with $s = 1$ as $s \to 1$. More precisely, the solution map

$$(\tau, \xi, s) \to X(\tau; \xi, s),$$

where $X(\tau; \xi, s)$ is the solution $X(\tau)$ of (4.28) with $X(0) = \xi$, is continuous on $\mathbb{R} \times \mathbb{R}^4 \times [1, \infty)$. In particular, this implies that the two periodic orbits Λ_- and Λ_+ depend continuously on s for $s \in [1, \infty)$.

In the limit case $s = 1$ the two periodic orbits on

$$T = \{(\xi_1, \xi_2, \xi_3, \xi_4) |\, \xi_1^2 + \xi_2^2 = \xi_3^2 + \xi_4^2 = C_0\}$$

are given by

$$\xi_1 \xi_3 - \xi_2 \xi_4 = 0, \tag{4.29}$$

or in terms of (4.22), by $\phi + \theta = \pm\frac{\pi}{2}$. This can be seen from a second conservation law that exists in the linear case: multiplying $u'''' + u = 0$ by u''' we have that $\frac{1}{2}u'''^2 + uu'' - \frac{1}{2}u'^2$ is constant. In particular, after transforming to the X-variables,

$$\frac{1}{2}X_4^2 + X_1 X_3 - \frac{1}{2}X_2^2 = 0$$

is invariant, whence (4.29), which defines two circles on the torus T.

4.4. Small solutions. We observed in section 4.1 that the role of $X_5 = 0$ and $X_5 = \infty$ may be reversed. This is a direct consequence of the scaling invariance of (4.1). Thus we may also use (4.4) for the analysis of small solutions to (4.1). The situation is depicted schematically in Figure 4.2. We simply apply (4.4) with a negative λ so that $X_5 \to 0$ corresponds to $u \to 0$. This changes only the sign in (4.10) for X_5 and means that the orbit Λ_+ now lies in the part of $X_5 = 0$ which repels solutions with $X_5 > 0$. Hence the stable manifold of Λ_+ is contained in $\Pi \cap \{X_5 = 0\}$. The unstable manifold of Λ_+ is given by the direct product $\Lambda_+ \times \{X_5 \,|\, X_5 > 0\}$ and has dimension 2. In the original variables it is the unstable manifold of $u = 0$ if $s = 1$ and the center-unstable manifold if $s > 1$. Likewise, the stable manifold of Λ_- is $\Lambda_- \times \{X_5 \,|\, X_5 > 0\}$, i.e., the direct product of Λ_- and the positive X_5-axis. As we have seen in section 4.3, the limit $s \to 1$ is well behaved in the X-variables.

We will use this analysis of the behavior near the equilibrium point $u = 0$ in section 5 to perform a continuous deformation of the stable manifold for $s = 1$ to the center-stable manifold for $s > 1$. We remark that, based on the similarity of the linear and nonlinear problem, the equilibrium point $u = 0$ of (4.1) for $s > 1$ can be considered as the nonlinear equivalent of a saddle-focus.

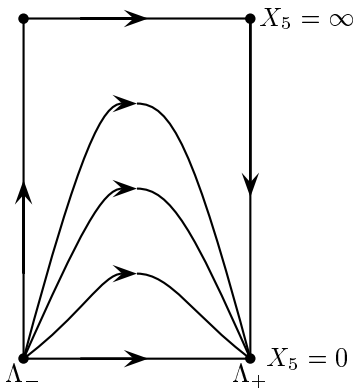


FIG. 4.2. *A schematic view of the flow on the Poincaré cylinder* Π *for the equation* $u'''' + u^s = 0$. *The role of* $X_5 = 0$ *and* $X_5 = \infty$ *is reversed when* λ *is negative.*

4.5. The full system. Applying the Poincaré transformation (4.4) with exponents (4.6) to the differential equation (1.3), or, more generally, to

$$x_1' = x_2; \quad x_2' = x_3; \quad x_3' = x_4; \quad x_4' = \Phi(x_1, x_2, x_3, x_4),$$

we arrive at

$$\dot{X}_1 = X_2 P - 4X_1 Q\,; \tag{4.30a}$$

$$\dot{X}_2 = X_3 P - (3+s) X_2 Q\,; \tag{4.30b}$$

$$\dot{X}_3 = X_4 P - (2+2s) X_3 Q\,; \tag{4.30c}$$

$$\dot{X}_4 = \Psi P - (1+3s) X_4 Q\,; \tag{4.30d}$$

$$\dot{X}_5 = -\frac{1}{\lambda} X_5 Q\,, \tag{4.30e}$$

where

$$Q = X_1^s X_2 + X_4 \Psi + X_3 (X_2 + X_4) \tag{4.31}$$

and

$$\Psi = X_5^{4\lambda s} \Phi \left(\frac{X_1}{X_5^{4\lambda}}, \frac{X_2}{X_5^{(3+s)\lambda}}, \frac{X_3}{X_5^{(2+2s)\lambda}}, \frac{X_4}{X_5^{(1+3s)\lambda}} \right). \tag{4.32}$$

In the case of (1.3) we have

$$\Phi(x_1, x_2, x_3) = \alpha x_3 - c x_2 + f(x_1),$$

where $f(x_1) = -x_1^3 + g(x_1)$ with $g(x_1)$ compactly supported. With $s = 3$ and $\lambda = \frac{1}{2}$ we thus obtain

$$\Psi = -X_1^3 + \alpha X_3 X_5^2 - c X_2 X_5^3 + g\left(\frac{X_1}{X_5^2}\right) X_5^6. \tag{4.33}$$

The last term in (4.33) is C^2 and has its derivatives up to second order vanishing in $X_5 = 0$. The extra terms are thus at least quadratic in X_5 for small X_5. Therefore

the local analysis near $X_5 = 0$ and in particular the Floquet multipliers of $\Lambda_\pm$ in the previous section are completely unaffected. The flow on the sphere Σ (at infinity) is identical to the flow for the reduced equation (4.2). Only the flow on $\Pi \setminus \Sigma$ is different. Note that in this analysis it is essential that the exponent s is larger than 1. We have the following theorem (compare Lemmas 4.3, 4.5 and 4.6).

THEOREM 4.8. *Let f satisfy hypothesis* (3.1) *and let $\alpha, c \in \mathbb{R}$.*

(a) *The stable periodic orbit Λ_+ of* (4.11) *is an asymptotically stable periodic orbit of* (4.30) *with nontrivial Floquet multipliers in* $(0,1)$. *Every solution of* (1.3) *which is unbounded in forward time corresponds to a solution of* (4.30) *having Λ_+ as ω-limit set. A similar statement holds for solutions unbounded in backward time and Λ_-.*

(b) *Unbounded solutions of* (1.3) *blow up oscillatorily in finite time.*

(c) *If $c \neq 0$, the energy $\mathcal{E}$ also blows up.*

Proof. By Lemma 4.6 all solutions of (4.30) which lie in the invariant set $\Pi \cap \{X_5 = 0\} \setminus \Lambda_- \subset \Pi$ tend to Λ_+ in forward time. Reminiscent of the proof of Lemma 4.6 we choose a small negatively invariant open tubular neighborhood Λ_-^ε of Λ_- in Π. By compactness of $\Pi \cap \{X_5 = 0\}$ there exists an open neighborhood Σ^ε of $\Pi \cap \{X_5 = 0\} \setminus \Lambda_-^\varepsilon$ in Π such that all orbits with starting point in Σ^ε tend to Λ_+ in forward time. Clearly every unbounded solution of (1.3) enters Σ^ε and thus tends to Λ_+.

For part (b) we observe that the exponent b in (4.8) is smaller than 1 so that in the old time variable X_5 can only go to zero in finite time. Finally we have that the energy $\mathcal{E}$ can remain bounded only if its derivative is integrable. For $c \neq 0$ this implies that u' is square integrable (see (1.5)) and thus u itself is (locally) bounded, which prohibits finite time blow-up, a contradiction. ∎

Remark 4.9. Theorem 4.8 establishes that large solutions of (1.3) are really described by oscillating solutions of $u'''' + u^3 = 0$. Thus large solutions do not "see" the other terms in (1.3) as they oscillate away to infinity. This is not only true for perturbations of the form $-u^3 + g(u)$ with g compactly supported and smooth, but also for global lower order perturbations. For such lower order perturbations Theorem 4.8 applies as well.

5. The winding number for small speeds. In this section we prove part (a) of Proposition 3.3. Before we can prove this theorem we first need a description of the global behavior of $W^s(1)$ for $c = 0$. In the following lemma we show that for $\alpha > \frac{1}{\sqrt{\sigma(f)}}$ all orbits in the stable manifold $W^s(1)$ are unbounded, and, after transforming to the X-variables in section 4, they all have Λ_- as α-limit set. Because all the nontrivial Floquet multipliers of Λ_- lie in $(1, \infty)$ (see Theorem 4.8(a)), this remains true for $c > 0$ sufficiently small.

LEMMA 5.1. *Let f satisfy hypothesis* (3.1), *let $\alpha > \frac{1}{\sqrt{\sigma(f)}}$, and let $c = 0$. Then $W^s(1)$ consists of unbounded orbits only, all of which connect Λ_- to $u = 1$.*

Proof. The proof is a combination of arguments also used in [25]. Any bounded solution must have its range in the set

$$V = \{u \in \mathbb{R} \mid F(u) \geq F(-a)\}$$

because a solution reaching outside this interval oscillates away toward infinity, as mentioned in Remark 4.7. Besides, any bounded solution must have at least one minimum below the line $u = -a$, again basically by the same oscillation argument as in the proof of Lemma 4.5. We now assume, arguing by contradiction, that u is a bounded orbit in $W^s(1)$. We will show that the range of u is not contained in V, so

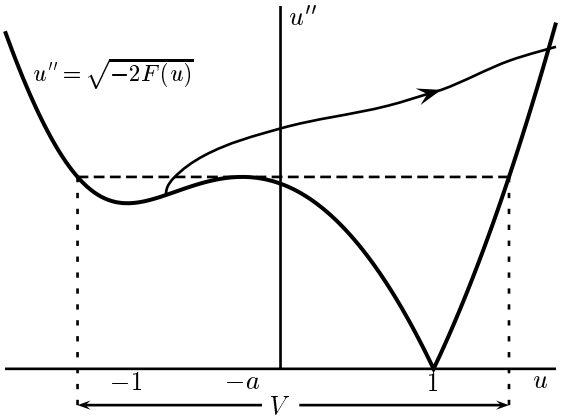


FIG. 5.1. *The (u, u'')-plane with the curve $u'' = \sqrt{-2F(u)}$. We have sketched the orbit of u for $t \geq t_0$, which is discussed in the proof of Lemma* 5.1. *We have also indicated the set V, in which every bounded solution has its range.*

that u is in fact unbounded. It then follows from Theorem 4.8 that u tends to Λ_- as $t \to -\infty$.

Thus, suppose that u is a bounded solution in $W^s(1)$. Changing from t to $-t$ if necessary we have that in such a minimum (using the fact that $\mathcal{E}(u, u', u'', u''') = 0$)

$$u(t_0) \leq -a, \quad u'(t_0) = 0, \quad u''(t_0) = \sqrt{-2F(u(t_0))} > 0, \quad u'''(t_0) \geq 0. \tag{5.1}$$

We will show that $u(t)$ increases to a value outside V for $t > t_0$, which immediately leads to a contradiction.

Define an auxiliary function

$$G(t) \stackrel{\text{def}}{=} u''(t) - \sqrt{-2F(u(t))}.$$

The following line of reasoning is depicted in Figure 5.1. First, $G(t_0) = 0$ and we show that $G(t) > 0$ in a right neighborhood of t_0. It is seen from the condition on α and the observation that $f(u) > 0$ on $(-\infty, -1) \cup (-a, 1)$ that

$$f(u) > -\sqrt{-\frac{\alpha^2}{2} F(u)} \qquad \text{for } u < 1. \tag{5.2}$$

If $u'''(t_0) > 0$, then clearly $G'(t_0) > 0$, whereas when $u'''(t_0) = 0$, then $G'(t_0) = 0$, and (since $u'(t_0) = 0$)

$$G''(t_0) = u''''(t_0) + \frac{f(u(t_0))}{\sqrt{-2F(u(t_0))}} u''(t_0) = \alpha \sqrt{-2F(u(t_0))} + 2f(u(t_0)) > 0$$

by the differential equation, and (5.1) and (5.2). Thus $G(t) > 0$ in a right neighborhood of t_0.

Second, we show that $G(t) > 0$ as long as $u(t) < 1$. We define $t_1 > t_0$ as the first maximum of $u(t)$ and $t_2 > t_0$ as the first point where $G(t_2) = 0$ (a priori, both t_1 and t_2 may be ∞). Then $t_2 < t_1$ since $u''(t) > 0$ as long as $G(t) > 0$. It now follows from the expression (1.4) for the energy and by (5.2) that

$$G'(t) = u'''(t) + \frac{f(u(t))}{\sqrt{-2F(u(t))}} u'(t)$$

$$= \frac{\frac{1}{2}u''^2(t) + F(u(t))}{u'(t)} + \left(\frac{\alpha}{2} + \frac{f(u(t))}{\sqrt{-2F(u(t))}}\right) u'(t)$$
$$> 0$$

as long as $G(t) > 0$ and $u(t) < 1$. Since $G(t) > 0$ in a right neighborhood of t_0 this implies that $G(t) > 0$ and $G'(t) > 0$ as long as $u(t) < 1$, and thus $u(t_2) \geq 1$.

Finally, we define $t_3 > t_0$ as the first point where $u(t) = -a$. It is easily seen that $t_3 < t_2$. By the energy expression we have that $u'''(t) > 0$ as long as $G(t) > 0$, thus $u''(t_2) > u''(t_3) > \sqrt{-2F(-a)}$. Combining the inequalities $u(t_2) \geq 1$ and $F(u(t_2)) = -\frac{1}{2}u''^2(t_2) < F(-a)$, we infer that $u(t_2)$ lies outside V, so that u is unbounded. By Theorem 4.8 all these unbounded orbits converge to Λ_-. □

Remark 5.2. Because all the nontrivial Floquet multipliers of Λ_- lie in $(1, \infty)$ (see Theorem 4.8(a)), Lemma 5.1 remains true for $c > 0$ sufficiently small.

The following theorem is equivalent to Proposition 3.3(a). We recall that K is defined in (3.3), and that its boundary δK is a level set of the energy.

THEOREM 5.3. *Let f satisfy hypothesis* (3.1) *and let $\alpha > \frac{1}{\sqrt{\sigma(f)}}$. For $F(-a) < E_0 < F(-1)$ let K be defined by* (3.3) *and let $W^s(1)$ be the stable manifold of the equilibrium $u = 1$. Then, provided $c > 0$ is sufficiently small, $W^s(1) \cap \delta K$ is a topological circle. Its projection Γ on the (u, u''')-plane winds exactly once around a disk containing both closed curves defined by $E_0 - F(u) + \frac{1}{2\alpha}u'''^2 = 0$ (see also Figure* 3.2*), i.e., $n(\Gamma, -1) = n(\Gamma, 1) = 1$.*

Proof. Our strategy is to deform $f(u)$ in several steps to the pure cubic $-u^3$ and let α go to zero. We have to do this in such a way that for each intermediate f the conclusion of Lemma 5.1 remains valid. All orbits in the stable manifold $W^s(1)$ thus tend to Λ_- in backward time, and this remains true during the entire deformation process. At the end of the deformation process we arrive at the reduced equation $u'''' + u^3 = 0$. We then use the analysis performed in section 4 to find a precise description of the orbits in $W^s(1)$. Finally, we obtain the results of Theorem 5.3 for the original equation (1.3) via continuation arguments.

Recall that $f(u) = -u^3 + g(u)$ with g having compact support, say, $g(u) = 0$ for all $|u| \geq C_0$. Taking C_0 sufficiently large, define the cut-off function $\phi \in C_0^\infty$ with $0 \leq \phi \leq 1$, $\phi(y) = 1$ for $|y| \leq C_0$, and $\phi(y) = 0$ for $|y| > C_0 + 1$.

Step 1. First deform $f(u)$ to a function which changes sign at $u = 1$ only. Let

$$f_\lambda(u) = f(u) - \lambda(u-1)\phi(u).$$

For λ large enough, say, $\lambda > \lambda_0$, the function $f_\lambda(u)$ has a zero at $u = 1$ only.

LEMMA 5.4. *Let $\alpha > \frac{1}{\sqrt{\sigma(f)}}$ and replace $f(u)$ by $f_\lambda(u)$. Then for all $\lambda \in [0, \lambda_0]$ the stable manifold $W^s(1)$ consists of unbounded orbits only, all of which connect Λ_- to $u = 1$.*

Proof. Let $\lambda_1 = \inf\{\lambda \mid f_\lambda(u) > 0 \text{ for all } u < 1\}$. For any $\lambda < \lambda_1$ the argument is exactly the same as in the proof of Lemma 5.1, where we use the following generalized definition of σ:

$$\sigma(f_\lambda) = \min\left\{ \frac{-F(u)}{2f(u)^2} \;\middle|\; u < 1 \text{ and } f(u) < 0 \right\}.$$

Note that $\sigma(f_\lambda) \leq \sigma(f_0)$ for $0 < \lambda < \lambda_1$ since $f_\lambda(u)$ and $-F_\lambda(u)$ are increasing in λ for all $u < 1$. For $\lambda \geq \lambda_1$ the result also holds, but by a different and less

restrictive oscillation argument, which applies to any $f(u)$ with a single zero at which it goes from positive to negative, and all $\alpha \geq 0$. We already used this in the proof of Lemma 4.6; the argument showing that every solution $u \not\equiv 1$ oscillates toward infinity is almost identical (for $\alpha \geq 0$ the second order term does not cause any difficulties). This completes the proof of the lemma. □

Continuing with the proof of Theorem 5.3, we change f to $f^1 \stackrel{\text{def}}{=} f_{\lambda_0}$ by letting λ go from 0 to λ_0. This leaves the local structure near $X_5 = 0$, and in particular near Λ_-, unaffected (see section 4.5).

Step 2. We change $f^1(u) = -u^3 + g^1(u)$ with $g^1(u) = g(u) - \lambda_0(u-1)\phi$ to $f^2(u) \stackrel{\text{def}}{=} -u^3(1-\phi) - (u-1)\phi$. Using the deformation functions

$$f_\lambda(u) = -u^3(1-\phi(u)) + (1-\lambda)(-u^3\phi(u) + g^1(u)) - \lambda(u-1)\phi(u),$$

we let λ go from 0 to 1, thus continuously deforming f^1 into f^2. All orbits in $W^s(1)$ are still unbounded and tend to Λ_- as $t \to -\infty$ during this deformation, since $f_\lambda(u)$ has a single zero at which it goes from positive to negative (see the proof of Lemma 5.4).

Step 3. It is now easy to shift the zero to the origin. Define

$$f_\lambda(u) = -u^3(1-\phi(u)) - (u - (1-\lambda))\phi(u).$$

Letting λ change from 0 to 1 deforms f^2 into $f^3 \stackrel{\text{def}}{=} -u^3(1-\phi) - u\phi$. Since we have shifted the origin we now have $W^s(0)$ instead of $W^s(1)$. All orbits in $W^s(0)$ are still unbounded and tend to Λ_- as $t \to -\infty$.

Step 4. Next we let α go to zero. The stable manifold $W^s(0)$ changes smoothly and the local structure near Λ_- again remains unaffected because α only appears in terms quadratic in X_5. For $\alpha = 0$ we have arrived at the equation

$$u'''' - f^3(u) = 0, \quad \text{with } f^3(u) = -u^3(1-\phi) - u\phi.$$

Step 5. We change f^3 using a family of functions

$$f_s(u) = -u^3(1-\phi) - u^s\phi.$$

Letting s increase from $s = 1$ to $s = 3$ we obtain a function $f^4(u) \stackrel{\text{def}}{=} u^3$. We note (see section 4.4) that for $s > 1$ the manifold W is the center-stable manifold of 0. Here we use section 4.3 to conclude that in this process W changes continuously, with the orbits in manifold $W = W^{cs}(0)$ still tending to Λ_- in backward time.

By sections 4.1 and 4.4 we have that, after going through Steps 1–5, W is the product of Λ_- and the X_5-axis. In view of the nontrivial Floquet multipliers of Λ_- being in $(1, \infty)$, it holds that for any small $\varepsilon > 0$ there exists a negatively invariant tubular neighborhood Λ_-^ε of Λ_- in Π with

$$\Lambda_-^\varepsilon \subset \{X = (X_1, X_2, X_3, X_4, X_5) \in \Pi \,|\, d(X, \Lambda_-) < \varepsilon\}.$$

We can choose this neighborhood such that

$$\overline{\Lambda_-^\varepsilon} \cap \{X_5 = \varepsilon\} \;=\; \{(X_1, X_2, X_3, X_4) \in \Lambda_-,\, X_5 = \varepsilon\}. \tag{5.3}$$

Besides, we can choose Λ_ε such that the flow for our final equation $u'''' + u^3 = 0$ is transversal to $\delta\Lambda_-^\varepsilon$. Moreover, for $\varepsilon > 0$ sufficiently small, we can choose Λ_ε such the flow is transversal to $\delta\Lambda_-^\varepsilon$ for every intermediate $f(u)$ and α in the deformation process of Steps 1–5 above, hence also for the original equation (1.3) with $c = 0$.

For any given $r > 0$ we can choose $\varepsilon > 0$ so small that the projection Γ_ε of $W \cap \delta\Lambda_-^\varepsilon$ on the (x_1, x_4)-plane (or, equivalently, on the (u, u''')-plane) is a curve with minimal distance to the origin at least r. To see this, we observe that the solution of (4.1) represented by Λ_- cannot have a point where $u = u''' = 0$, for in such a point also $u'' = 0$ in view of the energy $\mathcal{E}$ being zero. This would contradict the fact that $Q < 0$ on Λ_-. Thus in the X-variables Λ_- is uniformly bounded away from $(X_1, X_4) = (0, 0)$, so that for any $r > 0$ we can find an $\varepsilon > 0$ such that the projection of Λ_-^ε on the (u, u''')-plane has a distance larger than r from the origin. Therefore, the winding numbers around $u = \pm 1$ of the projection Γ_ε of $W \cap \delta\Lambda_-^\varepsilon$ on the (u, u''')-plane are well defined for ε sufficiently small.

It follows from (5.3) that for our final equation $u'''' + u^3 = 0$ we have

$$W \cap \delta\Lambda_-^\varepsilon = \{(X_1, X_2, X_3, X_4, X_5) \,|\, (X_1, X_2, X_3, X_4) \in \Lambda_-,\ X_5 = \varepsilon\},$$

so that, choosing r large, $n(\Gamma_\varepsilon, -1) = n(\Gamma_\varepsilon, 1) = 1$. By continuity the winding numbers of Γ_ε do not change if we reverse Steps 1–5, and again by continuity arguments and Remark 5.2 this remains true for $c > 0$ sufficiently small.

Finally, for our original equation (1.3) we know that, tracing back orbits in $W^s(1)$ until they hit $\delta\Lambda_-^\varepsilon$, their energy $\mathcal{E}$ remains close to 0, provided we keep $c > 0$ sufficiently small. Thus $W^s(1) \cap \delta K$ is contained in Λ_-^ε for small $c > 0$. Following $W^s(1) \cap \delta\Lambda_-^\varepsilon$ backwards along the flow to $W^s(1) \cap \delta K$ (which is a transversal intersection for $c > 0$), we see that the winding numbers $n(\Gamma, \pm 1)$ of the projection of $W^s(1) \cap \delta K$ are also 1. This completes the proof of Theorem 5.3. □

6. The winding number for large speeds. In this section we proof part (b) of Proposition 3.3.

THEOREM 6.1. *Let f satisfy hypothesis* (3.1) *and let $\alpha \in \mathbb{R}$. For $c > 0$ sufficiently large the intersection of the stable manifold $W^s(1)$ of $u = 1$ and the boundary δK of K is a smooth simple closed curve which projects on a closed curve Γ in the (u, z)-plane with $n(\Gamma, -1) = 0$ and $n(\Gamma, 1) = 1$.*

Proof. We first prove the theorem for a deformation of $f(u)$. We choose the nonlinearity $\tilde{f}(u)$ to satisfy

$$\tilde{f}(u) = f'(1)(u - 1) \quad \text{in a neighborhood } B_\varepsilon(1) \text{ of } u = 1.$$

For this deformed nonlinearity $\tilde{f}$ we compute the energy $\tilde{\mathcal{E}}$ on a closed curve in $\tilde{W} = W^s(1)$ winding once around $u = 1$ with u-values contained in $B_\varepsilon(1)$. The equation is now linear near $u = 1$, and the characteristic equation

$$-\mu^4 + \alpha\mu^2 + f'(1) = c\mu$$

has two eigenvalues $-\mu_1$ and $-\mu_2$ with negative real part (recall that $f'(1) < 0$). For $c > 0$ large enough μ_1 and μ_2 are real, and asymptotically

$$\mu_1 \sim c^{\frac{1}{3}} \quad \text{and} \quad \mu_2 \sim \frac{-f'(1)}{c} \qquad \text{as } c \to \infty. \tag{6.1}$$

Since the equation is linear, $\tilde{W}$ is given by (for c large enough)

$$\tilde{W} = \{(u, v, w, z) \,|\, u = u(t) = 1 + A_1 e^{-\mu_1 t} + A_2 e^{-\mu_2 t},\ v = u'(t),\ w = u''(t),\ z = u'''(t)\}. \tag{6.2}$$

We may choose a curve $S_1 \subset \tilde{W}$ around $u = 1$ parametrized by $\phi \in [0, 2\pi)$ by taking $t = 0$ and $A_1 = r\cos\phi$, $A_2 = r\sin\phi$ in (6.2) for some fixed $r > 0$. The projection of S_1 on the (u, u''')-plane is given by

$$\{(u, z) \,|\, u = 1 + r(\cos\phi + \sin\phi),\ z = -r(\mu_1^3\cos\phi + \mu_2^3\sin\phi),\ 0 \le \phi < 2\pi\}.$$

The energy on S_1 is given by

$$\begin{aligned} -\mathcal{E} &= \int_0^\infty cu'(t)^2 dt = c\int_0^\infty (A_1\mu_1 e^{-\mu_1 t} + A_2\mu_2 e^{-\mu_2 t})^2 dt \\ &= c\left(\frac{A_1^2\mu_1}{2} + \frac{2A_1A_2\mu_1\mu_2}{\mu_1+\mu_2} + \frac{A_2^2\mu_2}{2}\right) = c\mu_2\left(\frac{A_1^2\mu_1}{2\mu_2} + \frac{2A_1A_2\mu_1}{\mu_1+\mu_2} + \frac{A_2^2}{2}\right). \end{aligned} \tag{6.3}$$

Using (6.1) and estimating (6.3) from below we have, for c sufficiently large,

$$\mathcal{E} \le \frac{f'(1)}{4} r^2 < 0 \ \text{ on } S_1.$$

Thus, choosing an energy level $0 > \tilde{E}_0 > \frac{f'(1)}{4} r^2$, we have that S_1 lies in the complement of K. Let $\tilde{S} = \tilde{W} \cap \delta\tilde{K}$. Then $\tilde{S}$ lies inside S_1 and is obtained by tracing solutions in (6.2) of the linear equation forwards in time until they enter $\tilde{K}$. It follows that $\tilde{S}$ winds around $u = 1$ in $\tilde{W}$ exactly once and therefore its projection $\tilde{\Gamma}$ on the (u, z)-plane winds once around $(u, z) = (1, 0)$.

The calculations above involve only u-values between $1 - r\sqrt{2}$ and $1 + r\sqrt{2}$ so we may change the definition of $\tilde{f}(u)$ outside this range. In particular, taking r small, we may choose $\tilde{f}(u)$ such that $\tilde{F}(u)$ has a minimum $\tilde{F}(-a) < \tilde{E}_0$ and a maximum $\tilde{F}(-1) \in (\tilde{E}_0, \tilde{F}(1))$, with $-1 < -a < 1 - r\sqrt{2}$. Clearly $\tilde{\Gamma}$ does not wind around the point $(u, z) = (-1, 0)$.

We continue $\tilde{f}$ to f and $\tilde{E}_0$ to E_0, taking c large enough as to stay within a class of nonlinearities for which there does not exist a connection between $u = -1$ and $u = 1$ (see Lemma 2.1). By continuity we still have that $n(\Gamma, -1) = 0$ and $n(\Gamma, 1) = 1$. □

7. Travelling waves connecting an unstable to a stable state. In this section we focus on travelling waves that connect the unstable state $u = -a$ to one of the two stable states $u = \pm 1$. As in the proof of Theorem 1.1 in section 3 we begin by reducing to nonlinearities f which satisfy (3.1).

To obtain the necessary bound for $\alpha > 0$ we fix $c > 0$ and simply follow the argument in the proof of Lemma 2.2 with $F(-1)$ replaced by $F(-a)$ (for connections from $-a$ to $+1$), or by $F(-1) - F(-a)$ (for connections from $-a$ to -1).

By different methods it is also possible to prove a priori bounds in the case that $\alpha \le 0$. Applying a result by Gallay [14] to the present context we obtain the following. Let f satisfy (H_1), i.e., $\lim_{|u|\to\infty} \frac{f(u)}{u} = -\infty$. Then for any $\alpha \in \mathbb{R}$ there exists a constant C_0 such that any travelling wave solution $u(t, x) = U(x + ct)$ of (1.1) satisfies $\|u\|_\infty \le C_0$. The constant C_0 will thus depend only on α and $m \overset{\text{def}}{=} \sup\{|u| : \frac{f(u)}{u} \ge -D_\alpha\}$, where $D_\alpha > 0$ is a constant which depends on α only.

The idea is to consider $\Phi_y(t) = \int_{-\infty}^\infty h_y(x)u^2(t, x)dx$, where $h_y(x) = \frac{1}{1+(x-y)^2}$. Using the differential equation (1.1) one obtains an estimate of the form $\frac{d\Phi_y}{dt} \le A_0 - \Phi_y$ for some constant A_0 independent of y and t, hence $\Phi_y(t) \le A_0 + \Phi_y(0)e^{-t}$. Defining $\Psi(t) = \sup_{y\in\mathbb{R}} \Phi_y(t)$ one derives that for travelling waves, Ψ is independent of t, hence

$\Psi \leq A_0$. Combining this with the fact that $\int_{-\infty}^{\infty}(\frac{du}{dx})^2 dx = \frac{F(\pm 1)-F(-a)}{c}$, one then obtains an L^∞-bound on u.

Thus, for every $c > 0$ there exists a constant $C_0 > 0$ such that any solution of (1.3) connecting $-a$ to ± 1 satisfies $\|u\| < C_0$. This a priori estimate implies that we may replace f by $\tilde{f}(u) = \phi(u)f(u) - u^3(1-\phi(u))$, where the cut-off function $\phi \in C_0^\infty$ is such that $0 \leq \phi \leq 1$, $\phi(y) = 1$ for $|y| \leq C_0$, and $\phi(y) = 0$ for $|y| > C_0 + 1$. As in section 3 it holds that u is a travelling wave solution with speed c for nonlinearity $f(u)$ if and only if u is a travelling wave solution with speed c for nonlinearity $\tilde{f}(u)$.

The above argument shows that, looking for travelling waves, we may as well assume that f satisfies (3.1). The next theorem thus proves Theorem 1.2.

THEOREM 7.1. *Let f satisfy hypothesis* (3.1) *and let $\alpha \in \mathbb{R}$. For every $c > 0$ there exists a solution of* (1.3) *connecting $u = -a$ to $u = -1$.*

Proof. For all $c > 0$ we have that the three equilibria are hyperbolic and

$$\dim W^s(\pm 1) = \dim W^u(\pm 1) = 2, \quad \dim W^u(-a) = 3, \quad \dim W^s(-a) = 1.$$

Travelling wave solutions connecting $u = -a$ and $u = -1$ correspond to a nonempty intersection of $W^u(-a)$ and $W^s(-1)$. Recall that

$$\mathcal{E}(u, u', u'', u''') = -u'u''' + \frac{1}{2}u''^2 + \frac{\alpha}{2}u'^2 + F(u), \quad \text{where} \quad F(u) = \int_1^u f(s)ds,$$

satisfies (1.5). We take $F(-1) < E_1 < F(1)$ and consider the set

$$\tilde{K} = \left\{ (u, v, w, z) \,|\, E(u, v, w, z) = -vz + \frac{1}{2}w^2 + \frac{\alpha}{2}v^2 + F(u) \leq E_1 \right\}.$$

Now suppose that for some $c > 0$ the theorem is false. Then all orbits in $W^u(-a)$ have to leave $\tilde{K}$ through $\delta\tilde{K}$, because an orbit with bounded energy has no other choice than to converge to an equilibrium (see the proof Lemma 3.2) and $u = -1$, the only equilibrium in $\tilde{K}$ with energy larger than $E(-a)$, is excluded by assumption. Thus we have that the intersection of $W^u(-a)$ and $\delta\tilde{K}$ is homeomorphic to a 2-sphere S^2.

For the moment we consider the case that $\alpha > 0$. Since $\delta\tilde{K}$ is given by

$$\alpha\left(v - \frac{z}{\alpha}\right)^2 + w^2 = 2E_1 - 2F(u) + \frac{z^2}{\alpha}, \tag{7.1}$$

we may deform it smoothly into

$$\{(u, v, w, z) \,|\, u^2 + z^2 = 1 + v^2 + w^2\},$$

which defines a 3-manifold homeomorphic to $\mathbb{R}^2 \times S^1$. As deformations we use

$$(\lambda\alpha + 1 - \lambda)\left(v - \lambda\frac{z}{\alpha}\right)^2 + w^2 = G(u, \lambda) + \left(1 - \lambda + \frac{\lambda}{\alpha}\right)z^2,$$

with λ running from 1 to 0, and $G(u, 1) = 2E_1 - 2F(u)$ and $G(u, 0) = -1 + u^2$. Singularities can only appear in points on these manifolds where $G_u = v = w = z = 0$ and can thus be avoided by the choice of E_1.

It follows that $\delta\tilde{K}$ is homeomorphic to $\mathbb{R}^2 \times S^1$, or, equivalently, to the open solid torus. The intersection $W^u(-a) \cap \delta\tilde{K}$, being homeomorphic to S^2, divides $\delta\tilde{K}$ into two components, one bounded and homeomorphic to an open ball in $\mathbb{R}^3$, the other

unbounded. This division is in fact not completely straightforward. One needs to lift (a neighborhood of) $W^u(-a) \cap \delta\tilde{K}$ to the universal covering space $\mathbb{R}^3$ of $\tilde{K}$ and show that the unbounded part of the complement of the countable union of lifts is path-connected. Using the fact that the intersection $W^u(-a) \cap \delta\tilde{K}$ is induced by a flow, one can invoke the generalized Schoenflies theorem (see [5, Theorem 19.11]) to conclude that a lift of $W^u(-a) \cap \delta\tilde{K}$ divides $\mathbb{R}^3$ into an unbounded and a bounded component, which is homeomorphic to an open ball, in $\mathbb{R}^3$. Besides, the bounded components of the countable infinity of lifts can be contracted to points. The unbounded component (the complement of the countable union of bounded components) is thus homeomorphic to $\mathbb{R}^3 \setminus \mathbb{Z}$, hence path-connected.[5]

Now consider the piecewise smooth 3-manifold formed by the disjoint union of the point $(-a, 0, 0, 0)$, $W^u(-a) \cap \tilde{K}$ and the bounded component of $\delta\tilde{K} \setminus (W^u(-a) \cap \delta\tilde{K})$. This 3-manifold is homeomorphic to two closed balls in $\mathbb{R}^3$ sharing an S^2, namely, $W^u(-a) \cap \delta\tilde{K}$, as boundary and is therefore homeomorphic to an S^3. By the Jordan–Brouwer theorem this 3-manifold divides $\mathbb{R}^4$ to two components, one bounded, the other unbounded. We notice that the bounded component is negatively invariant. Clearly both components contain exactly one of the two orbits which together form $W^s(-a)$. Now consider the orbit in $W^s(-a)$ contained in the bounded component (which is negatively invariant). Since its energy is bounded we may, again by the argument in the proof of Lemma 3.2, conclude that, tracing it backwards, it must go to an equilibrium with energy less than the energy of $u = -a$. Since such an equilibrium does not exist, we have arrived at a contradiction.

The cases $\alpha < 0$ and $\alpha = 0$ are similar, the only changes being that we deform $\delta\tilde{K}$, given by (7.1), to $u^2 + v^2 = 1 + z^2 + w^2$ if $\alpha < 0$, and that for $\alpha = 0$ we rewrite $\delta\tilde{K}$ as $-2vz + w^2 = 2E_1 - 2F(u)$, which deforms into $-2vz + w^2 = -1 + u^2$ or $\frac{1}{2}(v+z)^2 + u^2 = \frac{1}{2}(v-z)^2 + w^2 + 1$. This completes the proof of the theorem. □

Remark 7.2. In the proof of Theorem 7.1 above we have used the nondegeneracy of the equilibrium point $u = -a$, while $u = -1$ may degenerate (i.e., $f'(-1) = 0$). The theorem also holds when $u = -a$ is degenerate but $u = -1$ is nondegenerate; in this case the argument in the proof of Theorem 7.3 below can be used. If $F(-1) = F(1)$ one also applies the proof of Theorem 7.3; see Remark 7.4.

Next we prove Theorem 1.3. Let

$$c^* \stackrel{\text{def}}{=} \inf\{\tilde{c} > 0 \,|\, \text{there is no connection from } -1 \text{ to } +1 \text{ for } c > \tilde{c}\}.$$

From Lemma 2.1 we see that c^* is well defined, and $c^* > 0$ for $\alpha > \frac{1}{\sqrt{\sigma(f)}}$ by Theorem 1.1. The argument at the beginning of this section shows that, in order to prove Theorem 1.3, we may restrict to nonlinearities f which satisfy (3.1). If $c_* > 0$, then it follows from Lemma 3.2 that for $c = c^*$ there exists a solution of (1.3) which connects -1 to $+1$. The following theorem thus proves both Theorem 1.3 and Corollary 1.4.

THEOREM 7.3. *Let f satisfy hypothesis* (3.1) *and let $\alpha \in \mathbb{R}$. For every $c > c^*$ there exists a solution of* (1.3) *connecting $u = -a$ to $u = 1$.*

Proof. We consider the stable manifold $W = W^s(1)$ of $u = 1$. We have shown in Theorem 6.1 that for $c > 0$ large enough the intersection of the stable manifold W of $u = -1$ and the boundary δK of K (defined in (3.3)) is a smooth simple closed

[5]We gratefully acknowledge several discussions with H. Geiges. He showed us that, via the Jordan–Brouwer separation theorem and an inductive Mayer–Vietoris argument, the division of $\delta\tilde{K}$ into two components can also be derived without using the extra information provided by the flow.

curve which projects on a closed curve Γ in the (u,z)-plane with $n(\Gamma,-1)=0$ and $n(\Gamma,1)=1$. It follows from the definition of c^* and Lemma 3.2 that, by continuity, this remains true for all $c>c_*$. Now fix $c>c^*$.

Let us assume by contradiction that there is no connection between $u=-a$ and $u=1$. The intersection between W and δK depends continuously on the energy level E as long as we do not encounter an equilibrium point. Assuming there is no connection between $u=-a$ and $u=1$, we let E decrease from $F(-1)>E_0>F(-a)$ to $E_2<F(-a)$. The projection Γ in the (u,z)-plane then depends continuously on E, as do the winding numbers, so that $n(\Gamma,-1)=0$ and $n(\Gamma,1)=1$ for all $E_0\le E\le E_2$. However, for the energy level E_2 we have that $(-1,0)$ and $(1,0)$ lie in the same component of the complement of the projection of δK onto the (u,z)-plane. Therefore $n(\Gamma,-1)=n(\Gamma,1)$, a contradiction. □

Remark 7.4. When $F(-1)=F(+1)$, then the same method shows that there exist travelling waves connecting $u=-a$ to $u=\pm1$ for all $c>0$ and all $\alpha\in\mathbb{R}$. Besides, as already noted in Remark 7.2, the method in the proof of Theorem 7.3 can be used to obtain an alternative proof of Theorem 7.1.

Finally, we prove Theorem 1.5 which deals with nonlinearities with two zeros (and a different behavior for $u\to\pm\infty$).

THEOREM 7.5. *Let $\alpha\in\mathbb{R}$ and let f satisfy hypothesis* (H_2). *For every $c>0$ there exists a solution of* (1.3) *connecting $u=0$ to $u=1$.*

Proof. Since the shape of the nonlinearity differs significantly from the one considered so far, we cannot invoke Lemma 3.2 directly. Besides, we find a priori bounds via a slightly different method.

Let $D\overset{\text{def}}{=}\sup\{\tilde u<0\,|\,F(u)>0 \text{ on } (-\infty,\tilde u)\}$. Travelling wave solutions connecting 0 to 1 satisfy $u\ge D$, since it follows from (1.4) and (1.5) that u can have no extremum in the range $u<D$ (at an extremum one would have $\mathcal{E}>F(1)$, which is impossible). Therefore, we may without loss of generality replace f by any function f_1 for which $f_1(u)=f(u)$ for $u\ge D$, and $f_1(u)<0$ for $u<D$. We choose f_1 such that $f_1(u)=u$ for $u<D-1$.

Now that we have a bound from below, we can also obtain a bound from above. A connecting solution of (1.3) is also a solution of (1.3) with f_1 replaced by any f_2 for which $f_2(u)=f_1(u)$ for all $u\ge D-1$. We choose $f_2(u)=-u^3$ for $u<D-2$, and argue as at the beginning of this section to conclude that there exists a uniform bound $\|u\|_\infty\le C_0$ on all travelling wave solutions. We may thus replace f_1 by a function f_3 for which $f_3(u)=f_1(u)$ for $u\le C_0$ and $f_3(u)=-u^3$ for $u\ge C_0+1$. We conclude that u is a travelling wave solution with speed c for nonlinearity $f(u)$ if and only if u is a travelling wave solution with speed c for nonlinearity $f_3(u)$.

In the following we therefore assume, without loss of generality, that $f(u)=u$ for $u\le D-1$, and $f(u)=-u^3$ for $u\ge C_0+1$.

We now follow the argument in the proof of Lemma 3.2. However, we cannot use Lemma 3.1 to show that orbits in $W^s(1)$, which are completely contained in K, are bounded. Instead, we argue as follows. Suppose, by contradiction, that an orbit $u(t)$ in $W^s(1)$ is completely contained in K and is unbounded. As in the proof of Lemma 3.2 it follows from (3.4) that $u(t)$ exists for all $t\in\mathbb{R}$. There are now two possibilities: either $u(t)\ge D-1$ for all $t\in\mathbb{R}$ or there exists some $t_0\in\mathbb{R}$ such that $u(t_0)<D-1$. First we deal with the latter case.

Since (see above) $u(t)$ cannot attain an extremum in the range $u<D$, it follows that $u(t)$ is increasing for $t\le t_0$. Hence $u(t)$ obeys, for $t\le t_0$, the linear equation $cu'=-u''''+\alpha u''+u$. Since u is unbounded as $t\to-\infty$, it follows that $u=$

$-a_0 e^{-a_1 t} + o(1)$ for some $a_0, a_1 > 0$ as $t \to -\infty$. By substituting this into (3.4) a contradiction is reached.

Next we deal with the case where $u(t) \geq D - 1$ for all $t \in \mathbb{R}$. Clearly $u(t)$ is a solution of (1.3) with f replaced by any function $\tilde{f}$ for which $\tilde{f}(u) = f(u)$ for all $u \geq D-1$. We choose $\tilde{f}(u) = -u^3$ for $u < D-2$, and it follows from Lemma 3.1 that u blows up in finite time, a contradiction.

Having circumvented the problem in the proof of Lemma 3.2 we conclude that for $F(0) < E_0 < F(-1)$ the intersection of the stable manifold W of $u = -1$ and the boundary δK of K (defined in (3.3)) is a smooth simple closed curve which projects on a closed curve Γ in the (u, z)-plane with $n(\Gamma, 1) = 1$.

The rest of the argument is analogous to the proof of Theorem 7.3. Assuming that there is no connection between $u = 0$ and $u = 1$, the final contradiction is now obtained by the fact that $n(\Gamma, 1) = 0$ for $E_2 < F(0)$. □

8. Concluding remarks. The most apparent open problem concerns the range of α-values for which a travelling wave connecting -1 to $+1$ exists. For some examples it can be shown that such a travelling wave does not exist for all $\alpha \in \mathbb{R}$. The more general question whether for any nonlinearity satisfying (H_1) a bound α_* exists such that there are no travelling waves for $\alpha < \alpha_*$ remains open.

Regarding the uniqueness of the various travelling wave solutions not much is known. For large α (i.e., $\gamma \approx 0$) the travelling wave connecting -1 to $+1$ may be expected to be unique (analogous to the limiting second order case). The results in [6] show that uniqueness does not hold for $f_a(u) = (u + a)(1 - u^2)$ with a small when $\alpha < \sqrt{8}$. Equation (1.1) with $f(u) = u - u^3$ admits an abundance of standing wave solutions for $0 \leq \alpha < \sqrt{8}$. It has been proved in [6] that these solutions can be perturbed to travelling waves for $f_a(u)$ with small a and small $c = c(a)$. Since this can be done for any standing wave, an infinite family of curves in the (a, c)-plane passing through the origin is thus obtained.

The method used in this paper does not give any information about the shape of the solution. For example, we would like to know for which values of α the solution is monotone. Since we do not know the value of c for which a traveling wave occurs, we do not in general even know whether the connected equilibrium points are approached monotonically or in an oscillatory manner.

Finally, the question arises to what extent the travelling wave solution is of importance to the dynamics of the PDE. It might be a limit profile for a broad class of initial conditions as is the case for the second order equation [13]. Since travelling waves connecting $u = -a$ to $u = \pm 1$ exist for large ranges of c, it would be interesting to know which of these waves is generally encountered. In [9, 12] the wave selection mechanism has been investigated for a propagating front which is formed from localized initial data (i.e., $u + a$ is localized). Using the physically motivated assumption that the linearized equation (around $u = -a$) drives the system, it is argued that for $\alpha > \sqrt{12 f'(-a)}$ one of the travelling waves is selected (and the wave speed is calculated), while for $\alpha < \sqrt{12 f'(-a)}$ the propagating front is argued not to have a fixed profile. However, the only rigorous stability result that we know of is of a perturbative nature [27] (i.e., α very large) and moreover it does not answer the question of the selection of the wave speed.

REFERENCES

[1] N.N. Akhmediev, A.V. Buryak, and M. Karlsson, *Radiationless optical solitons with oscil-*

lating tails, Opt. Commun., 110 (1994), pp. 540–544.
[2] M.E. Akveld and J. Hulshof, *Travelling wave solutions of a fourth-order semilinear diffusion equation*, Appl. Math. Lett., 11 (1998), pp. 115–120.
[3] C.J. Amick and J.F. Toland, *Homoclinic orbits in the dynamic phase space analogy of an elastic strut*, European J. Appl. Math., 3 (1992), pp. 97–114.
[4] D.G. Aronson and H.F. Weinberger, *Nonlinear diffusion in population genetics, combustion, and nerve pulse propagation*, in Partial Differential Equations and Related Topics, J.A. Goldstein, ed., Lecture Notes in Math. 446, Springer-Verlag, Berlin, 1975.
[5] G.E. Bredon, *Topology and Geometry*, Springer-Verlag, New York, 1993.
[6] B. Buffoni, *Shooting methods and topological transversality*, Proc. Roy. Soc. Edinburgh Sect. A, 129 (1999), pp. 1137–1155.
[7] B. Buffoni, A.R. Champneys, and J.F. Toland, *Bifurcation and coalescence of multi-modal homoclinic orbits for a Hamiltonian system*, J. Dynam. Differential Equations, 8 (1996), pp. 221–281.
[8] C. Conley, *Isolated invariant sets and the Morse index*, CBMS Reg. Conf. Ser. Math. 38, AMS, Providence, RI, 1978.
[9] G.T. Dee and W. van Saarloos, *Bistable systems with propagating fronts leading to pattern formation*, Phys. Rev. Lett., 60 (1988), pp. 2641–2644.
[10] F. Dumortier, *Singularities of vector fields on the plane*, J. Differential Equations, 23 (1977), pp. 53–106.
[11] F. Dumortier, *Local study of planar vector fields: Singularities and their unfoldings*, in Structures in Dynamics, Finite Dimensional Deterministic Studies, H.W. Broer et al., eds., Stud. Math. Phys., Vol. 2, North-Holland, Amsterdam, 1991, pp. 161–241.
[12] U. Ebert and W. van Saarloos, *Universal algebraic relaxation of fronts propagating into an unstable state and implications for moving boundary approximations*, Phys. Rev. Lett., 80 (1998), pp. 1650–1653.
[13] P.C. Fife and J.B. McLeod, *The approach of solutions of nonlinear diffusion equations to travelling front solutions*, Arch. Rational Mech. Anal., 65 (1977), pp. 335–361.
[14] T. Gallay, *private communication*.
[15] R.A. Gardner and C.K.R.T. Jones, *Traveling waves of a perturbed diffusion equation arising in a phase field model*, Indiana Univ. Math. J., 39 (1990), pp. 1197–1222.
[16] W.D. Kalies, J. Kwapisz, J.B. van den Berg, and R.C.A.M. Vandervorst, *Homotopy classes for stable periodic and chaotic patterns in fourth-order Hamiltonian systems*, Comm. Math. Phys., 214 (2000), pp. 573–592.
[17] W.D. Kalies, J. Kwapisz, and R.C.A.M. Vandervorst, *Homotopy classes for stable connections between Hamiltonian saddle-focus equilibria*, Comm. Math. Phys., 193 (1998), pp. 337–371.
[18] W.D. Kalies and R.C.A.M. Vandervorst, *Multitransition homoclinic and heteroclinic solutions to the extended Fisher-Kolmogorov equation*, J. Differential Equations, 131 (1996), pp. 209–228.
[19] J. Kwapisz, *Uniqueness of the stationary wave for the extended Fisher-Kolmogorov equation*, J. Differential Equations, 165 (2000), pp. 235–253.
[20] P.J. Olver, *Applications of Lie Groups to Differential Equations*, Springer-Verlag, New York, 1993.
[21] G. Palis, *Linearly induced vector fields and $\mathbb{R}^2$-actions on spheres*, J. Differential Equations, 23 (1977), pp. 53–106.
[22] L.A. Peletier, A.I. Rotariu-Bruma, and W.C. Troy, *Pulse-like spatial patterns described by higher-order model equations*, J. Differential Equations, 150 (1998), pp. 124–187.
[23] L.A. Peletier and W.C. Troy, *Spatial patterns described by the extended Fisher-Kolmogorov (EFK) equation: Kinks*, Differential Integral Equations, 8 (1995), pp. 1279–1304.
[24] L.A. Peletier and W.C. Troy, *A topological shooting method and the existence of kinks of the extended Fisher-Kolmogorov equation*, Topol. Methods Nonlinear Anal., 6 (1995), pp. 331–355.
[25] M.A. Peletier, *Non-existence and uniqueness results for fourth-order Hamiltonian systems*, Nonlinearity, 12 (1999), pp. 1555–1570.
[26] L. Perko, *Differential Equations and Dynamical Systems*, Springer-Verlag, New York, 1991.
[27] V. Rottschäfer and C.E. Wayne, *Existence and stability of traveling fronts in the extended Fisher-Kolmogorov equation*, J. Differential Equations, to appear.
[28] J.B. van den Berg, *Uniqueness of solutions for the extended Fisher-Kolmogorov equation*, C. R. Acad. Sci. Paris Sér. I Math., 326 (1998), pp. 447–452.
[29] J.B. van den Berg, *The phase-plane picture for some fourth-order conservative differential equations*, J. Differential Equations, 161 (2000), pp. 110–153.

SIAM J. MATH. ANAL.
Vol. 32, No. 6, pp. 1375–1389

ISOSPECTRAL FLOWS OF THIRD ORDER OPERATORS*

L. AMOUR†

Abstract. We consider the third order linear differential operator $L_{p,q} = i\frac{d^3}{dx^3} + i\frac{d}{dx}q + iq\frac{d}{dx} + p$ on the unit interval. The associated boundary conditions are $y(0) = y(1) = 0$, $y'(0) = zy'(1)$ with $z \in \mathbf{C}$, $|z| = 1$. For all but a finite number of parameters z, the eigenvalues λ_j of $L_{p,q}$ are of multiplicity one. For fixed boundary conditions we make the formulae for isospectral flows induced by the Hamiltonian function λ_j explicit. These flows commute.

Key words. inverse spectral theory, third order operator, isospectral flows

AMS subject classifications. 34A55, 34L05, 47E05

PII. S0036141098347249

1. Introduction and results.

In this paper we consider the third order linear differential operator

$$L_{p,q} = i\frac{d^3}{dx^3} + i\frac{d}{dx}q + iq\frac{d}{dx} + p \tag{1}$$

on the unit interval. The potentials p and q are real valued functions. It is supposed that $(p,q) \in L^2_R \times H^1_R$. We will often use the abbreviated notation $' = \partial/\partial x$.

Let us introduce the boundary conditions depending on a real parameter $\phi \in T = R|_{2\pi Z}$:

$$(BC)_\phi \qquad \begin{cases} y(0) = y(1) = 0, \\ y'(0) = e^{i\phi}y'(1). \end{cases} \tag{2}$$

Fix ϕ in T. The operator $L_{p,q}$ associated with the boundary conditions $(BC)_\phi$ is self-adjoint in $L^2_{\mathbf{C}}[0,1]$ with the scalar product $(f,g) = \int_0^1 f\bar{g}\,dx$ and has a discrete spectrum. Each eigenvalue is real and of multiplicity one or two. Let $(\lambda_j(p,q,\phi))_{j\in\mathbf{Z}}$ denote the increasing sequence of eigenvalues.

The main result is Theorem 3. Preliminary results on the spectrum of $L_{p,q}$ are given in Theorems 1 and 2. Note that the description of the spectrum is not as precise as the one in [A2]. In particular, there is no counting lemma here.

THEOREM 1. *Fix $(p,q) \in L^2_R \times H^1_R$. For all but a finite number of parameters $\phi \in T$ the eigenvalues of $L_{p,q}$ associated with the boundary conditions $(BC)_\phi$ are of multiplicity one.*

From now on, therefore, it is possible to restrict ourselves to simple spectra. The sequence $(\lambda_j(p,q,\phi))_{j\in\mathbf{Z}}$ denotes a strictly increasing sequence of eigenvalues. Let $h_j(x,p,q,\phi)$ be the eigenfunctions corresponding to $\lambda_j(p,q,\phi)$ with $\|h_j(\cdot,p,q,\phi)\|_{L^2_{\mathbf{R}}[0,1])} = 1$. Thus the function h_j is determined up to the phase. This phase is irrelevant in the following. Set $\langle q\rangle = \int_0^1 q(x)\,dx$.

THEOREM 2. *Suppose $(p,q,\phi) \in L^2_R \times H^1_R \times T$. Then*

$$\text{(i)}\quad \lambda_j(p,q,\phi) = (2j\pi)^3 + (2j\pi)^2(3\phi - \pi) + 2j\pi\left(3\left(\phi - \frac{\pi}{3}\right)^2 - 2\langle q\rangle\right) + O(1)$$

*Received by the editors November 9, 1998; accepted for publication (in revised form) August 11, 2000; published electronically March 28, 2001.
http://www.siam.org/journals/sima/32-6/34724.html

†Laboratoire de Mathématiques, UPRESA 6056, Université de Reims, Moulin de la Housse, B.P. 1039, 51687 Reims Cedex 2, France (laurent.amour@univ-reims.fr).

as $j \to +\infty$, *and*

$$\text{(ii)} \quad \lambda_j(p,q,\phi) = (2j\pi)^3 + (2j\pi)^2(3\phi+\pi) + 2j\pi\left(3\left(\phi+\frac{\pi}{3}\right)^2 - 2\langle q\rangle\right) + O(1)$$

as $j \to -\infty$.

In particular, $\langle q\rangle$ and ϕ are spectral invariants. We introduce the Poisson bracket

$$\{F,G\} = \int_0^1 \frac{\partial F}{\partial p}\left(\frac{\partial G}{\partial q}\right)' - \frac{\partial G}{\partial p}\left(\frac{\partial F}{\partial q}\right)' dx$$

for smooth real valued functions of (p,q). It is skew-symmetric and satisfies Jacobi's identity. In other words, the space of potentials is equipped with the symplectic structure $\omega(\mathbf{D}\xi,\mathbf{D}\eta) = \int_0^1 \xi_1\eta_2' - \xi_2'\eta_1\,dx$, where the skew-symmetric operator $\mathbf{D} = \begin{pmatrix} 0 & \frac{d}{dx} \\ \frac{d}{dx} & 0 \end{pmatrix}$.

Set $N_j = \begin{pmatrix} \frac{\partial\lambda_j}{\partial p} \\ \frac{\partial\lambda_j}{\partial q} \end{pmatrix}$. For $F = \lambda_j$, $G = \lambda_k$ one has

$$\{\lambda_j,\lambda_k\} = \int_0^1 N_j^\top \mathbf{D} N_k\,dx.$$

Set $V_k = \mathbf{D}N_k$. The bracket $\{\lambda_j,\lambda_k\}$ is, according to the chain rule, the change of λ_j under the Hamiltonian flow induced by the vector field V_k. We shall prove that $\{\lambda_j,\lambda_k\} = 0\,\forall\, j,k$. This means that all λ_j are integrals of the motion induced by the vector field V_k. Then the corresponding flow is isospectral. The flows induced by the vectors fields V_j and V_k commute. This is also a consequence of $\{\lambda_j,\lambda_k\} = 0$. The existence of these isospectral flows is given by explicit formulae. Our main result is the following.

THEOREM 3. *The solution to*

$$\frac{d}{dt}\begin{pmatrix} p(\cdot,t) \\ q(\cdot,t) \end{pmatrix} = V_j(\cdot, p(t), q(t), \phi)$$

with initial value (p_0,q_0) *with* $h_j'(0,p_0,q_0) = 0$ *exists for every* $t \in \mathbf{R}$ *and is given by*

$$p(x,t) = p_0(x) + 3(e^t-1)\frac{d}{dx}\frac{\operatorname{Im}\left(h_j(x,p_0,q_0)\bar{h}_j'(x,p_0,q_0)\right)}{\theta_j(x,t,p_0,q_0)}, \tag{3}$$

$$q(x,t) = q_0(x) + \frac{3}{2}\frac{d^2}{dx^2}\log\theta_j(x,t,p_0,q_0), \tag{4}$$

where

$$\theta_j(x,t,p_0,q_0) = 1 + (e^t-1)\int_0^x |h_j(s,p_0,q_0)|^2\,ds.$$

Theorem 3 proves that some isospectral flows, in the third order case and for fixed boundary conditions, are given by a very particular explicit formula of the following form: *initial potential* $+ G$, where G depends only on t and on one eigenfunction of the initial potential.

The proof of Theorem 3 is then carried out in a constructive way. We determine a function $G = G(t, h_j)$ *together with* the associated eigenfunctions $h_j(t)$ requiring that the flow *initial potential* $+ G$ is isospectral while knowing the expression of its associated vectors field V_j in term of $h_j(t)$. These two requests lead to a unique G. However, when the initial eigenfunction verifies $h_j'(0) \neq 0$, the formulae (3) and (4) are not valid. The reason may be one of the following two: the form *initial potential* $+ G$ is too restrictive, or the boundary conditions $(BC)_\phi$ have to be generalized by introducing one or more parameters which should change with t. Nevertheless, one should note that for third order operators and contrarily to second order ones (see [IT]) there always exist (see next paragraph) some isospectral flows in the form *initial potential* $+ G$ for various parameters ϕ and for *fixed* boundary conditions.

The assumption $h_j'(0, p_0, q_0) = 0$ in Theorem 3 is satisfied for a large class of potential (p_0, q_0). Let us give examples. Set $r_0 \in H_R^1[0,1]$ such that the Schrödinger operator $-\frac{d^2}{dx^2} + r_0(x)$ has the origin as eigenvalue with the Dirichlet boundary conditions on the unit interval, and let $y(x)$ be the associated eigenfunction. This is not restrictive since the translation of τ on the spectrum acts on the potential as $r_0 \to r_0 + \tau$. Choose ϕ_0 with Theorem 1 such that $L_{p_0=0, q_0=-2r_0}$ has a simple spectrum. Using $L_{0,-2r_0} y^2 = 2iy(y'' - r_0 y)' + 6iy'(y'' - r_0 y)$ one obtains that y^2 is an eigenfunction of $L_{0,-2r_0}$ and satisfies $(y^2)'(0) = 0$.

If the condition $h_j'(0, p_0, q_0) = 0$ was not imposed, there would be an explicit isospectral flow for each j. The number of *explicit* isospectral flows is then restricted. When $h_j'(0, p_0, q_0) \neq 0$, nonexplicit isospectral flows could exist.

Furthermore, the condition $h_j'(0, p_0, q_0) = 0$ may be not artificial since it appears again with a different method for deriving isospectral flows. Namely, the commutation method leads to Theorem 3 in the case where $p_0 = 0$ and when h_j corresponds to the eigenvalue $\mu_j = 0$ (or, equivalently, $p_0 \equiv \mu_j$). In particular, one recovers the formula (4), and (3) becomes $p(x,t) \equiv 0$. Let us briefly outline the steps. Set $D = \frac{1}{i}\frac{d}{dx}$, $A_h^+ = hD\frac{1}{h}$, and $A_h^- = \frac{1}{h}Dh$. Let $L_{p,q} h_k = \mu_k h_k \ \forall k$. Fix j and suppose first that $\frac{1}{2}(h_j')^2 - h_j h_j'' - q h_j^2 = 0$. Then it is easy to check that $A_{h_j}^- D A_{h_j}^+ = L_{p,q} - \mu_j$. Therefore, $\forall k \neq j$, $h_k^+ = D A_{h_j}^+ h_k$ and $h_j^+ = \frac{1}{h_j}(c_1(t) + c_2(t)\int_0^x h_j^2(s)ds + c_3(t)\int_0^x h_j^2(s)\int_0^s h_j(\sigma)d\sigma)$ satisfy $D A_{h_j}^+ A_{h_j}^- h_k^+ = (\mu_k - \mu_j) h_k^+ \ \forall k$. In these conditions, $D A_{h_j}^+ A_{h_j}^- = D A_{h_j^+}^- A_{h_j^+}^+$ and commuting again we find that $\forall k \neq j$, $\tilde{h}_k = A_{h_j^+}^+ h_k^+$ and $\tilde{h}_j = \frac{1}{h_j^+}$ satisfy $A_{h_j^+}^+ D A_{h_j^+}^- \tilde{h}_k = (\mu_k - \mu_j)\tilde{h}_k \ \forall k$. This proves that $A_{h_j^+}^+ D A_{h_j^+}^- + \mu_j$ is isospectral to $A_{h_j}^- D A_{h_j}^+ + \mu_j = L_{p,q}$. Turning back to the condition $\frac{1}{2}(h_j')^2 - h_j h_j'' - q h_j^2 = 0$, we have $h_j L_{0,q} h_j = D(\frac{1}{2}(h_j')^2 - h_j h_j'' - q h_j^2)$ so that $L_{0,q} h_j = \mu_j h_j$ and (2) implies $\frac{1}{2}(h_j')^2 - h_j h_j'' - q h_j^2 = \frac{1}{2}(h_j'(0))^2 - \mu_j \int_0^x h_j^2(s)ds$ which vanish if $\mu_j = 0$ and $h_j'(0) = 0$. The formulae resulting from $A_{h_j}^- D A_{h_j}^+$ are (3) with $p(x,t) \equiv 0$ and (4). The jth eigenfunction $\tilde{h}_j$ is the same as the one in section 3.

In the second order case, isospectral flows have been largely studied in various situations. For the Schrödinger operator $-\frac{d^2}{dx^2} + q(x)$, $x \in [0,1]$, $q \in L^2$ explicit formulae for isospectral flows are established for the Dirichlet boundary conditions in [PT], for the separated boundary conditions in [IT], for the periodic case in [FIT], and for the generalized periodic case in [RT]. Explicit formulae are derived in [GR] for the singular Schrödinger operator $-\frac{d^2}{dx^2} + \frac{2}{x^2} + q(x)$, $x \in [0,1]$, $q \in L^2$ with Dirichlet boundary conditions and in [CM] for the Dirichlet problem $\frac{d}{dx}(p^2 \frac{d}{dx} y) + \lambda p^2 y = 0$, $y(0) = y(1) = 0$. For each eigenvalue λ_j the explicit formula arises naturally as Hamiltonian phase flows

with the Hamiltonian function λ_j.[1] The corresponding explicit formulae are derived using commutation methods based on the Crum–Darboux method (see [PT, section 5], [GST, appendices A, B]). These Hamiltonian flows are decisive for a full description of isospectral sets. For unbounded domains, explicit formulae are obtained in [MT] for the perturbed quantum harmonic oscillator $-\frac{d^2}{dx^2}+x^2+q(x)$, $x\in\mathbf{R}$, $q\in S$ and recently for a general situation $-\frac{d^2}{dx^2}+q(x)$, $x\in\mathbf{R}$, $q\in L^2$ in [GST]. In [A1] and [AG] explicit formulae of isospectral flows are given for a first order differential operator. To the best of our knowledge, Schrödinger operators, first order systems, and now third order operators (1) are the only examples where explicit formulae are derived for the isospectral flows induced by the Hamiltonian functions λ_j.

Let us mention that in the Lax framework and in the periodic case, one step in the study of a wave equation is to introduce an auxiliary spectra (made up of simple eigenvalues), of an ordinary differential operator L, self-adjoint in most cases. The eigenvalues of L and the canonically conjugated coordinates constitute a coordinate system on the phase space for the evolution equation. The boundary conditions for L are the Dirichlet boundary conditions or are closely related. The isospectral flows are used to determine the exact image of the phase space by the coordinate system. The wave equation associated to the ordinary differential operator $L=L_{p,q}$ is the Boussinesq equation $\frac{\partial^2 q}{\partial t^2}=\frac{\partial^2}{\partial x^2}(\frac{4}{3}q^2+\frac{1}{3}\frac{\partial^2 q}{\partial x^2})$ (see [M] for more considerations). The boundary conditions in (2) are chosen so as to be close to the Dirichlet boundary conditions, to obtain a self-adjoint operator when associated with $L_{p,q}$, and to potentially give simple eigenvalues. The fact that this would provide an auxiliary spectra is beyond our purpose.

Sections 2, 3, and 4 are concerned with the proofs of Theorem 1, Theorem 2, and Theorem 3, respectively.

2. Proof of Theorem 1. This section is devoted to the proof of Theorem 1. The proof is a consequence of Lemmas 2 and 3. Lemma 2 follows from Theorem 4 and Lemma 1.

Consider $(p,q)\in L^2_R\times H^1_R$, and let $\lambda\in\mathbf{C}$. The functions $y_1(x,\lambda,p,q)$, $y_2(x,\lambda,p,q)$, $y_3(x,\lambda,p,q)$ are defined as the unique solutions to $L_{p,q}y(x)=\lambda y(x)$ for a.e. $x\in[0,1]$, satisfying the initial conditions

$$\begin{pmatrix} y_1(0,\lambda,p,q) & y_2(0,\lambda,p,q) & y_3(0,\lambda,p,q)\\ y_1'(0,\lambda,p,q) & y_2'(0,\lambda,p,q) & y_3'(0,\lambda,p,q)\\ y_1''(0,\lambda,p,q) & y_2''(0,\lambda,p,q) & y_3''(0,\lambda,p,q)\end{pmatrix}=Identity.$$

Every solution to $L_{p,q}y=\lambda y$ can be expressed as

$$y(x)=a\,y_1(x,\lambda,p,q)+b\,y_2(x,\lambda,p,q)+c\,y_3(x,\lambda,p,q), \tag{5}$$

where $(a,b,c)=(y(0),y'(0),y''(0))$.

THEOREM 4. *Fix $(p,q,\phi)\in L^2_R\times H^1_R\times T$.*

(i) *The eigenvalues of $L_{p,q}$ associated with the boundary conditions $(BC)_\phi$ are of multiplicity one or two. They are the roots of the function defined from $\mathbf{C}$ into $\mathbf{R}$*

$$\lambda\mapsto\Delta(\lambda,p,q,\phi)=\cos\left(\frac{\phi}{2}\right)\operatorname{Re}y_3(1,\lambda,p,q)-\sin\left(\frac{\phi}{2}\right)\operatorname{Im}y_3(1,\lambda,p,q). \tag{6}$$

[1] Other types of isospectral flows are induced by the Korteweg–deVries (KdV) hierarchy of Hamiltonian functions.

(ii) *Moreover, λ is an eigenvalue and of multiplicity two if and only if $\mathcal{M}(\lambda,p,q,\phi)=\left(\begin{smallmatrix}0&0\\0&0\end{smallmatrix}\right)$, where*

$$\mathcal{M}(\lambda,p,q,\phi)=\begin{pmatrix} y_2(1,\lambda,p,q) & y_3(1,\lambda,p,q)\\ y_2'(1,\lambda,p,q)-e^{i\phi} & y_3'(1,\lambda,p,q)\end{pmatrix}. \tag{7}$$

Proof of Theorem 4. Using (5) the boundary conditions $(BC)_\phi$ are

$$\mathcal{M}(\lambda,p,q,\phi)\begin{pmatrix} b\\ c\end{pmatrix}=\begin{pmatrix}0\\0\end{pmatrix},\quad (b,c)\neq(0,0).$$

Consequently, λ is an eigenvalue if and only if $\operatorname{Ker}\mathcal{M}(\lambda,p,q,\phi)\le 1$. Thus the eigenvalues are the roots of $\det\mathcal{M}(\lambda,p,q,\phi)=0$. Using $\bar{y}_3=y_3'y_2-y_3y_2'$ (see [A2, Lemma 4]) we have

$$\det\mathcal{M}(\lambda,p,q,\phi)=\bar{y_3}(1,\lambda,p,q)+e^{i\phi}y_3(1,\lambda,p,q). \tag{8}$$

From (8) it is easy to check that

$$\det\mathcal{M}(\lambda,p,q,\phi)=2e^{i\frac{\phi}{2}}\left(\cos\left(\frac{\phi}{2}\right)\operatorname{Re}y_3(1,\lambda,p,q)-\sin\left(\frac{\phi}{2}\right)\operatorname{Im}y_3(1,\lambda,p,q)\right).$$

This proves (i). The eigenvalue λ is of multiplicity two if and only if $\operatorname{Ker}\mathcal{M}(\lambda,p,q,\phi)=2$. This proves (ii). □

LEMMA 1. *Suppose $(p,q)\in L^2_R\times H^1_R$ and $\lambda\in\mathbf{R}$. Then*

$$y_3(1,\lambda,p,q)=0\Rightarrow y_3'(1,\lambda,p,q)=y_2(1,\lambda,p,q)=|y_2'(1,\lambda,p,q)|-1=0.$$

Proof of Lemma 1. We have (see [A2, Lemma 4])

$$\bar{y}_2(x,\lambda,p,q)=y_3'(x,\lambda,p,q)y_1(x,\lambda,p,q)-y_3(x,\lambda,p,q)y_1'(x,\lambda,p,q) \tag{9}$$

and

$$\bar{y}_3(x,\lambda,p,q)=y_3'(x,\lambda,p,q)y_2(x,\lambda,p,q)-y_3(x,\lambda,p,q)y_2'(x,\lambda,p,q) \tag{10}$$

for $(x,\lambda,p,q)\in[0,1]\times\mathbf{R}\times L^2_R\times H^1_R$. We omit (p,q) from the notations for brevity.

Using (10), $y_3(1,\lambda)=0\Rightarrow y_2(1,\lambda)=0$ or $y_3'(1,\lambda)=0$. From (9), $y_3(1,\lambda)=y_3'(1,\lambda)=0\Rightarrow y_2(1,\lambda)=0$. Consequently, we have

$$y_3(1,\lambda)=0\Rightarrow y_2(1,\lambda)=0. \tag{11}$$

Moreover, (10) gives $\bar{y_3'}(1,\lambda)=y_3''(1,\lambda)y_2(1,\lambda)-y_3(1,\lambda)y_2''(1,\lambda)$. Then with (11) we get

$$y_3(1,\lambda)=0\Rightarrow y_2(1,\lambda)=y_3'(1,\lambda)=0. \tag{12}$$

There is no second order derivative in $L_{p,q}$. Then the Wronskian

$$W(x,\lambda)=\begin{vmatrix} y_1(x,\lambda) & y_2(x,\lambda) & y_3(x,\lambda)\\ y_1'(x,\lambda) & y_2'(x,\lambda) & y_3'(x,\lambda)\\ y_1''(x,\lambda) & y_2''(x,\lambda) & y_3''(x,\lambda)\end{vmatrix}$$

does not depend on x and therefore is equal to 1.

Suppose $y_3(1,\lambda) = y_3'(1,\lambda) = y_2(1,\lambda) = 0$. From $W(1,\lambda) = 1$ we get

$$y_1(1,\lambda)y_2'(1,\lambda)y_3''(1,\lambda) = 1. \tag{13}$$

Differentiate (9) with respect to x to obtain

$$\bar{y}_2'(1,\lambda)y_3''(1,\lambda) = y_3''(1,\lambda)y_1(1,\lambda). \tag{14}$$

Equalities (13) and (14) show that $|y_2'(1,\lambda)| = 1$.

$$y_3(1,\lambda) = y_3'(1,\lambda) = y_2(1,\lambda) = 0 \Rightarrow |y_2'(1,\lambda)| = 1. \tag{15}$$

Therefore, (12) and (16) complete the proof. □

LEMMA 2. *Suppose $(p,q) \in L^2_R \times H^1_R$ and $\lambda \in \mathbf{R}$. Then $y_3(1,\lambda,p,q) = 0$ if and only if there exists a unique $\phi \in T$ such that λ is an eigenvalue of multiplicity two for the operator $L_{p,q}$ with the boundary conditions $(BC)_\phi$.*

Proof of Lemma 2. The proof is a direct consequence of Theorem 4 and Lemma 1. Indeed if $y_3(1,\lambda,p,q) = 0$, then from Lemma 1, there exists a unique $\phi \in T$ such that $y_3'(1,\lambda,p,q) = y_2(1,\lambda,p,q) = 0$ and $y_2'(1,\lambda,p,q) = e^{i\phi}$. Then from Theorem 4(ii), λ is a double eigenvalue for $(BC)_\phi$. Conversely, using Theorem 4(ii) again, $y_3(1,\lambda,p,q) = 0$ when λ is a double eigenvalue. □

LEMMA 3. *Suppose $(p,q) \in L^2_R \times H^1_R$. The function defined from $\mathbf{R}$ into $\mathbf{C}$*

$$X \mapsto y_3(1,X,p,q)$$

has no roots in $|X| \geq 54^3 e^{3\|(p,q)\|}$, where $\|(p,q)\| = \|p\|_{L^2_R[0,1]} + 2\|q\|_{L^2_R[0,1]} + \|q'\|_{L^2_R[0,1]}$.

Proof of Lemma 3. Set $\Xi(x,\lambda) = e^{(|\operatorname{Im} k| + |\operatorname{Im} \omega k| + |\operatorname{Im} \omega^2 k|)\frac{x}{2}}$, where $\omega = e^{2i\frac{\pi}{3}}$ and $k^3 = \lambda$. From [A2, Theorem 3] we have

$$|y_3(1,\lambda,p,q) - y_3(1,\lambda,0,0)| \leq \frac{3}{|\lambda|} e^{\|(p,q)\|} \Xi(1,\lambda) \tag{16}$$

for every $\lambda \in \mathbf{C}$. Let $k = X + iY$ with $(X,Y) \in \mathbf{R}^2$, and define the sector

$$\mathcal{S} = \left\{ k \in \mathbf{C};\ X > 0 \text{ and } |Y| < \sqrt{3}X \right\}.$$

From (16) we obtain

$$|y_3(1,\lambda,p,q) - y_3(1,\lambda,0,0)| \leq \frac{3}{|\lambda|} e^{\|(p,q)\| + \frac{\sqrt{3}}{2}X + \frac{|Y|}{2}} \tag{17}$$

for every λ with $k \in \mathcal{S}$.

Moreover, $y_3(x,\lambda,0,0) = -\frac{1}{3k^2}(e^{ikx} + \omega e^{iwkx} + \omega^2 e^{iw^2 kx})$. In the half plane $\{X \geq 0\}$ the main term of $y_3(x,\lambda,0,0)$ is $\omega^2 e^{iw^2 kx}$. Therefore, it is easy to obtain

$$|y_3(1,\lambda,0,0)| \geq \frac{e^{\frac{\sqrt{3}}{2}X + \frac{Y}{2}}}{3|k|^2} \left(1 - e^{-\sqrt{3}X} - e^{-\frac{\sqrt{3}}{2}X - \frac{3}{2}Y}\right) \tag{18}$$

in $\{X \geq 0\}$. In addition, if $Y \geq -1$ and $\frac{\sqrt{3}}{2}X \geq 2\log 2 + \frac{3}{2}$, then $e^{-\frac{\sqrt{3}}{2}X - \frac{3}{2}Y} \leq \frac{1}{4}$ and $e^{-\sqrt{3}X} \leq \frac{1}{4}$. Under these conditions (18) gives

$$|y_3(1,\lambda,0,0| \geq \frac{1}{6|k|^2} e^{\frac{\sqrt{3}}{2}X + \frac{Y}{2}}. \tag{19}$$

Combining (17) and (19) we deduce that

$$|y_3(1,\lambda,p,q)-y_3(1,\lambda,0,0)|<|y_3(1,\lambda,0,0)| \tag{20}$$

for any λ in $\mathcal{D}$, where

$$\mathcal{D}=\Big\{\lambda=k^3\in\mathbf{C};\ k\in\mathcal{S},\ \operatorname{Re}k>\frac{4\sqrt{3}}{3}\log 2+\sqrt{3},$$

$$\operatorname{Im}k>-1,\ |k|>18e^{\|(p,q)\|}\Big\}.$$

Since $y_3(1,\cdot,p,q)$ is an entire function (cf. [A2, Theorem 4]), Rouché's theorem is applied with (20) on any contours inside $\mathcal{D}$. Consequently, $y_3(1,\lambda,p,q)$ and $y_3(1,\lambda,0,0)$ have the same number of real roots greater than $18^3e^{3\|(p,q)\|}$. Using (19), $y_3(1,\lambda,0,0)$ has no root in $\mathcal{D}$. Thus $y_3(1,\lambda,p,q)$ has no real root greater than $18^3e^{3\|(p,q)\|}$. It can be proved similarly that $y_3(1,\lambda,p,q)$ has no real root smaller than $-18^3e^{3\|(p,q)\|}$. □

The proof of Theorem 1 is now a consequence of Lemmas 2 and 3.

Proof of Theorem 1. Fix (p,q) in $L^2_R\times H^1_R$. From Lemma 3 the roots of $\lambda\mapsto y_3(1,\lambda,p,q)$ belong to $\{\mu\in\mathbf{R};\ |\mu|<18^3e^{3\|(p,q)\|}\}$. By the analyticity properties of $y_3(1,\cdot,p,q)$ (see [A2, Theorem 4][2]) there is a finite number of roots of $y_3(1,\lambda,p,q)$ in $\{\mu\in\mathbf{R};\ |\mu|<18^3e^{3\|(p,q)\|}\}$. The roots of $y_3(1,\cdot,p,q)$ are necessarily real. Let $\nu_1,\nu_2,\dots,\nu_N$ be these points and using Lemma 1 we define the real numbers $\psi_1,\psi_2,\dots,\psi_N$ by $y_2'(1,\nu_j,p,q)=e^{i\psi_j}$, $j=1,2,\dots,N$.

Following Theorem 4(i) the ν_j's are eigenvalues of $L_{p,q}$ with the boundary conditions $(BC)_\phi$ and for any $\phi\in T$. In other words, the roots of $y_3(1,\cdot,p,q)$ are the particular eigenvalues that remain fixed as the boundary conditions change. The multiplicity of the ν_j's is one or two. Following Lemma 2 the ν_j's are of multiplicity one when $\phi\neq\psi_1,\psi_2,\dots,\psi_N$.

From Lemma 3 and Theorem 1(ii) all eigenvalues in $\{\mu\in\mathbf{R};\ |\mu|\geq 18^3e^{3\|(p,q)\|}\}$ are simple for the boundary conditions $(BC)_\phi$ with any $\phi\in T$. The proof is now complete. □

3. Proof of Theorem 2. Let us recall here that the eigenvalues $\lambda_j(p,q,\phi)$ are the roots of $\operatorname{Re}(e^{i\frac{\phi}{2}}y_3(1,\lambda,p,q))$ and are of multiplicity one for $|j|$ sufficiently large. Following Picard's iteration for the integral equation satisfied by $y_3(x,\lambda,p,q)$ (see [A2, section 1]) we obtain for all $\lambda\in\mathbf{C}$

$$y_3(1,\lambda,p,q)=y_3(1,\lambda,0,0)+\sum_{n\geq 1}c_n(1,\lambda,p,q), \tag{21}$$

where

$$\begin{aligned}c_0(x,\lambda,p,q)&=y_3(x,\lambda,0,0),\\ c_n(x,\lambda,p,q)&=\int_0^x\Big(-2y_3'(x-t,\lambda,0,0)q(t)+y_3(x-t,\lambda,0,0)(q'(t)\\ &\qquad +ip(t)\Big)c_{n-1}(t,\lambda,p,q)\,dt.\end{aligned} \tag{22}$$

[2]The hypothesis $q(0)=0$ in [A2, Theorem 4] is not necessary for y_3.

For all $(x,\lambda)\in[0,1]\times\mathbf{C}$

$$\begin{aligned} y_1(x,\lambda,0,0) &= \frac{1}{3}\left(e^{ikx}+e^{i\omega kx}+e^{i\omega^2 kx}\right),\\ y_2(x,\lambda,0,0) &= \frac{1}{3ik}\left(e^{ikx}+\omega^2 e^{i\omega kx}+\omega e^{i\omega^2 kx}\right),\\ y_3(x,\lambda,0,0) &= \frac{1}{3(ik)^2}\left(e^{ikx}+\omega e^{i\omega kx}+\omega^2 e^{i\omega^2 kx}\right), \end{aligned} \tag{23}$$

and

$$|y_j(x,\lambda,0,0)|\le\frac{3}{|k|^{j-1}}\Xi(x,\lambda), \tag{24}$$

where $\Xi(x,\lambda)=e^{(|\mathrm{Im}\frac{k}{2}|+|\mathrm{Im}\frac{\omega k}{2}|+|\mathrm{Im}\frac{\omega^2 k}{2}|)x}$. Therefore, (21), (22), (23), and (24) give

$$y_3(1,\lambda,p,q)=y_3(1,\lambda,0,0)+c_1(1,\lambda,p,q)+O\left(\frac{\Xi(1,\lambda)}{|k|^4}\right). \tag{25}$$

Using (22) with (24)

$$c_1(1,\lambda,p,q)=-2\int_0^1 y_3'(1-t,\lambda,0,0)y_3(t,\lambda,0,0)q(t)\,dt+O\left(\frac{\Xi(1,\lambda)}{|k|^4}\right). \tag{26}$$

The key formula is

$$\begin{aligned} y_3'(x-t,\lambda,0,0)y_3(x,\lambda,0,0) = \frac{i}{3\lambda}\Big(&y_1(x,\lambda,0,0)\\ &+\omega y_1(x+t(\omega-1),\lambda,0,0)+\omega^2 y_1(x+t(\omega^2-1),\lambda,0,0)\Big) \end{aligned} \tag{27}$$

for all $(x,\lambda)\in[0,1]\times\mathbf{C}$. This is easily checked using (23). For $\lambda\in\mathbf{R}$, $\Xi(1,\lambda)=O(e^{\frac{\sqrt{3}}{2}|k|})$, and from (26) and (27) one gets

$$\begin{aligned} c_1(1,\lambda,p,q) = -\frac{2i}{3\lambda}\Big(&y_{\ 1}(1,\lambda,0,0)\langle q\rangle+\omega I_1(\lambda,p,q)\\ &+\omega^2 I_2(\lambda,p,q)\Big)+O\left(\frac{e^{\frac{\sqrt{3}}{2}|k|}}{|k|^4}\right), \end{aligned} \tag{28}$$

where

$$I_l(\lambda,p,q)=\int_0^1 y_1(1+t(\omega^l-1),\lambda,0,0)q(t)\,dt,\quad l=1,2, \tag{29}$$

for all $\lambda\in\mathbf{R}$. From (23) we deduce

$$I_1(\lambda,p,q)=\frac{1}{3}\sum_{j=1}^{3}J_j(\lambda,p,q),$$

where

$$J_j(\lambda,p,q)=\int_0^1 e^{ik\omega^{j-1}(1-t)+ik\omega^j t}q(t)\,dt,\quad j=1,2,3, \tag{30}$$

for all $\lambda \in \mathbf{R}$. Integrating by parts J_j, $j = 1, 2, 3$, we deduce that

$$I_1(\lambda, p, q) = O\left(\frac{e^{\frac{\sqrt{3}}{2}|k|}}{|k|}\right). \tag{31}$$

Similarly, (31) is valid replacing $I_1(\lambda, p, q)$ by $I_2(\lambda, p, q)$. Consequently, (25) and (28) give for every $\lambda \in \mathbf{R}$

$$y_3(1, \lambda, p, q) = y_3(1, \lambda, 0, 0) - \frac{2i\langle q\rangle}{3\lambda} y_1(1, \lambda, 0, 0) + O\left(\frac{e^{\frac{\sqrt{3}}{2}|k|}}{|k|^4}\right). \tag{32}$$

Set $k_j^3 = \lambda_j(p, q, \phi)$. From (32) and $\operatorname{Re}\left(e^{i\frac{\phi}{2}} y_3(1, \lambda_j(p, q, \phi), p, q)\right) = 0$ we obtain

$$\operatorname{Re}\left(e^{i\frac{\phi}{2}} y_3(1, \lambda_j(p, q, \phi), 0, 0)\right) + \frac{2\langle q\rangle}{3\lambda_j} \operatorname{Re}\left(e^{i\left(\frac{\phi-\pi}{2}\right)} y_1(1, \lambda_j(p, q, \phi), 0, 0)\right)$$

$$= O\left(\frac{e^{\frac{\sqrt{3}}{2}|k_j|}}{|k_j|^4}\right). \tag{33}$$

Using (33) and (23)

$$\frac{2\langle q\rangle}{3\lambda} \operatorname{Re}\left(e^{i\left(\frac{\phi-\pi}{2}\right)} y_1(1, \lambda, 0, 0)\right) = \frac{2\langle q\rangle}{9\lambda}\left(e^{\frac{\sqrt{3}}{2}|k|} \cos\left(\frac{k - \phi + \pi}{2}\right) + O(1)\right) \tag{34}$$

as $0 \le k \to +\infty$. For $k_j \ge 0$ we set

$$k_j = 2j\pi + \phi - \frac{\pi}{3} + 2\delta_j, \tag{35}$$

where $|\delta_j| < \pi$. Then combining (33) with (34) we deduce that

$$\cos\left(\frac{\pi}{2} + \delta_j\right) = \frac{2}{3k_j}\langle q\rangle \cos\left(\frac{\pi}{3} + \delta_j\right) + O\left(\frac{1}{k_j^2}\right)$$

for $k_j \ge 0$. Consequently, $\delta_j \to 0$ as $j \to +\infty$. Then $\delta_j = -\frac{\langle q\rangle}{6j\pi} + O(\frac{1}{j^2})$ as $j \to +\infty$ and $k_j = 2j\pi + \phi - \frac{\pi}{3} - \frac{\langle q\rangle}{3j\pi} + O(\frac{1}{j^2})$ as $j \to +\infty$. This proves (i). The asymptotic (ii) $k_j = 2j\pi + \phi + \frac{\pi}{3} - \frac{\langle q\rangle}{3j\pi} + O(\frac{1}{j^2})$ as $j \to -\infty$ can be obtained similarly or by observing that λ is an eigenvalue for $L_{p,q}$ with the boundary conditions $(BC)_\phi$ if and only if $-\lambda$ is an eigenvalue for L_{p^*,q^*} with the boundary conditions $(BC)_{-\phi}$, where $p^*(x) = p(x)$ for every x in $[0, 1]$ and $q^*(x) = q(x)$ for a.e. x in $[0, 1]$. □

4. Proof of Theorem 3. Fix $(p, q) \in L^2_R \times H^1_R$. It is proved in section 2 that for all but a finite number of parameters $\phi \in T$ the eigenvalues of $L_{p,q}$ with boundary conditions $(BC)_\phi$ are of multiplicity one. In section 2 we consider such ϕ and omit it from the notations. By the simplicity of the $\lambda_j(p, q, \phi)$ and applying the implicit functions theorem with $\Delta(\lambda_j(p, q, \phi), p, q, \phi) = 0$, we obtain that each λ_j is a smooth function of (p, q). Then the computation of the $L^2_R[0, 1] \times L^2_R[0, 1]$ gradient $\nabla_{p(x),q(x)}\lambda_j(p, q, \phi)$ is given by Lemma 5. The scalar product associated with $L^2_R[0, 1] \times L^2_R[0, 1]$ is

$$\left\langle \begin{pmatrix} Y(x) \\ Z(x) \end{pmatrix}, \begin{pmatrix} \tilde{Y}(x) \\ \tilde{Z}(x) \end{pmatrix} \right\rangle = \int_0^1 Y(x)\tilde{Y}(x) + Z(x)\tilde{Z}(x)\, dx.$$

We shall prove that the flow $(p(t), q(t))$ associated with the vector field

$$\begin{pmatrix} 0 & \frac{d}{dx} \\ \frac{d}{dx} & 0 \end{pmatrix} \nabla_{p(x),q(x)} \lambda_j(p,q)$$

is isospectral. Moreover, when $h_j'(0,p,q) = 0$, this flow exists $\forall t \in \mathbf{R}$ and is given by (3) and (4). Define $[y,z](x) = y'(x)z(x) - y(x)z'(x)\ \forall x \in [0,1]$.

LEMMA 4. *Suppose* $(p,q) \in L^2_R \times H^1_R$. *For every* $j \in \mathbf{Z}$

$$\nabla_{p(x),q(x)} \lambda_j(p,q) = \begin{pmatrix} |h_j(x,p,q)|^2 \\ i[h_j, \bar{h}_j](x,p,q) \end{pmatrix} \forall\, x \in [0,1].$$

Proof of Lemma 4. For $p, q, q' \in C^0_R[0,1]$, h_j''' is continuous. Then

$$L_{p,q}\,(d_p h_j(u)) + u h_j = \lambda_j d_p h_j(u) + d_p \lambda_j(u) h_j \tag{36}$$

and

$$L_{p,q}\,(d_q h_j(v)) + i(v h_j)' + i v h_j' = \lambda_j d_q h_j(v) + d_q \lambda_j(v) h_j. \tag{37}$$

Since $L_{p,q}$ is self-adjoint, $\|h_j\| = 1$, and $h_j(0) = h_j(1) = 0$, we obtain from (36) and (37) that $d_p \lambda_j(u) = (|h_j|^2, u)_{L^2_C[0,1]}$ and $d_q \lambda_j(v) = (ih_j' \bar{h}_j - i h_j \bar{h}_j', u)_{L^2_C[0,1]}$. Therefore,

$$d_{p,q}\lambda_j(u,v) = \left\langle \begin{pmatrix} |h_j(x,p,q)|^2 \\ i[h_j, \bar{h}_j](x,p,q) \end{pmatrix}, \begin{pmatrix} u \\ v \end{pmatrix} \right\rangle_{L^2_R[0,1] \times L^2_R[0,1]}.$$

Then for $p, q, q' \in C^0_R[0,1]$ we have $\nabla_{p,q}\lambda_j = \begin{pmatrix} |h_j|^2 \\ i[h_j, \bar{h}_j] \end{pmatrix}$ and since $\nabla_{p,q}\lambda_j$, h_j and h_j' are continuous functions of (p,q) in $L^2_R \times H^1_R$, it is true for $(p,q) \in L^2_R \times H^1_R$. □

LEMMA 5. *For every* $(j,k) \in \mathbf{Z}^2$ *we have* $\{\lambda_j, \lambda_k\} = 0$.

Consequently, every λ_k is an integral of $\begin{pmatrix} \dot{p} \\ \dot{q} \end{pmatrix} = V_j(p,q)$ for fixed boundary conditions.

Proof of Lemma 5. The proof is similar to [M, section 6]. For every $(j,k) \in \mathbf{Z}^2$ we have

$$\begin{aligned} &\int_0^1 \begin{pmatrix} |h_j|^2 \\ i[h_j, \bar{h}_j] \end{pmatrix} \begin{pmatrix} 0 & \frac{d}{dx} \\ \frac{d}{dx} & 0 \end{pmatrix} \begin{pmatrix} |h_k|^2 \\ i[h_k, \bar{h}_k] \end{pmatrix} dx \\ &\quad = -\frac{i}{3} \int_0^1 2[h_j, \bar{h}_j]' |h_k|^2 - 2[h_k, \bar{h}_k]' |h_j|^2 \\ &\qquad + [h_k, \bar{h}_k](|h_j|^2)' - [h_k, \bar{h}_k](|h_j|^2)'\, dx \\ &\quad = \frac{2i}{3} \int_0^1 w(\bar{h}_j, h_k) h_j \bar{h}_k - w(\bar{h}_k, h_j) h_k \bar{h}_j\, dx, \end{aligned} \tag{38}$$

where $w(\bar{f}, g) = i(\bar{f}''g - \bar{f}'g' + \bar{f}g'' + 2q\bar{f}g)$. If $j = k$ the expression (38) vanishes. Moreover, it is easy to check that

$$\frac{d}{dx} w(\bar{f}, g) = (\lambda - \mu)\bar{f}g \tag{39}$$

when $L_{p,q} f = \lambda f$ and $L_{p,q} g = \mu g$. Then (39) and (38) give

$$\{\lambda_j, \lambda_k\} = w(\bar{h}_j, h_k) w(\bar{h}_k, h_j)|_{x=0}^{x=1},$$

which completes the proof with (2). □

We now prove Theorem 3.

Proof of Theorem 3. (p_0, q_0, ϕ_0) are fixed. Let $h_j = h_j(x, p_0, q_0, \phi_0)$ and $\lambda_j = \lambda_j(x, p_0, q_0, \phi_0)$. Observing Lemma 5, the proof is carried out by determining not only the potentials but also one eigenfunction. We look for $(h(x,t), p(x,t), q(x,t))$ satisfying

$$\|h(t)\|_{L^2_R[0,1]} = 1, \tag{40}$$

$$L_{p(t),q(t)}h(t) = \lambda_j h(t), \tag{41}$$

$$h(0,t) = h(1,t) = 0, \quad h'(0,t) = e^{i\phi}h'(1,t), \tag{42}$$

and

$$\frac{d}{dt}\begin{pmatrix} p(t) \\ q(t) \end{pmatrix} = \frac{3}{2}\begin{pmatrix} 0 & \frac{d}{dx} \\ \frac{d}{dx} & 0 \end{pmatrix}\begin{pmatrix} |h(t)|^2 \\ i[h(t), \bar{h}(t)] \end{pmatrix}. \tag{43}$$

The $(h(t), p(t), q(t))$ are expressed as

$$h(x,t) = b(t)\frac{h_j(x)}{\theta(x,t)} \tag{44}$$

and

$$p(x,t) = p_0(x) + A(x,t), \quad q(x,t) = q_0(x) + B(x,t), \tag{45}$$

where $b(t)$, $\theta(x,t)$, $A(x,t)$, and $B(x,t)$ are real valued functions. In (44) $b(t)$ is the L^2 unit normalization (40). Using

$$[h(t), \bar{h}(t)] = \frac{b^2(t)}{\theta^2(x,t)}[h_j(x), \bar{h}_j(x)]$$

and since $\frac{h_j}{\theta(t)}$ satisfies (42) under the assumption $h_j'(0) = 0$, we shall determine $b(t)$, $\theta(x,t)$, $A(x,t)$, and $B(x,t)$ satisfying

$$\left\| b(t)\frac{h_j}{\theta(t)} \right\|_{L^2_R[0,1]} = 1, \tag{46}$$

$$L_{p_0+A(t),q_0+B(t)}\left(\frac{h_j}{\theta(t)}\right) = \lambda_j \frac{h_j}{\theta(t)}, \tag{47}$$

and

$$\frac{d}{dt}\begin{pmatrix} A(t) \\ B(t) \end{pmatrix} = \frac{3}{2}b^2(t)\frac{d}{dx}\frac{1}{\theta^2(t)}\begin{pmatrix} i[h_j, \bar{h}_j] \\ |h_j|^2 \end{pmatrix}. \tag{48}$$

First we deal with (47). Combining

$$i\left(\frac{h_j}{\theta}\right)''' + 2i(q_0 + B)\left(\frac{h_j}{\theta}\right)' + (iq_0' + iB' + p + A)\left(\frac{h_j}{\theta}\right) = \lambda_j\left(\frac{h_j}{\theta}\right)$$

with $ih_j''' + 2iq_0h_j' + (iq_0' + p_0)h_j = \lambda_j h_j$ while using $(\frac{h_j}{\theta})''' = h_j'''\left(\frac{1}{\theta}\right) + 3h_j''\left(\frac{1}{\theta}\right)' + 3h_j'\left(\frac{1}{\theta}\right)'' + h_j\left(\frac{1}{\theta}\right)'''$ gives

$$\begin{aligned}(A+iB')\left(\frac{h_j}{\theta}\right) + 2iB\left(\frac{h_j}{\theta}\right)' + 2iq_0h_j\left(\frac{1}{\theta}\right)' \\ + 3ih_j''\left(\frac{1}{\theta}\right)' + 3ih_j'\left(\frac{1}{\theta}\right)'' + ih_j\left(\frac{1}{\theta}\right)''' = 0.\end{aligned} \tag{49}$$

The real part of (49) multiplied by $2\bar{h}_j$ is

$$2|h_j|^2A\left(\frac{1}{\theta}\right) + 2i[h_j,\bar{h}_j]B\left(\frac{1}{\theta}\right) + 3i[h_j,\bar{h}_j]'\left(\frac{1}{\theta}\right)' + 3i[h_j,\bar{h}_j]\left(\frac{1}{\theta}\right)'' = 0 \tag{50}$$

and the imaginary part of (49) multiplied by $2\bar{h}_j$ gives

$$\begin{aligned}2B'|h_j|^2\left(\frac{1}{\theta}\right) + 2B\left(\left(|h_j|^2\right)'\left(\frac{1}{\theta}\right) + 4B|h_j|^2\left(\frac{1}{\theta}\right)'\right) \\ + 4q_0|h_j|^2\left(\frac{1}{\theta}\right)' + 3(h_j''\bar{h}_j + h_j\bar{h}_j'')\left(\frac{1}{\theta}\right)' \\ + 3\left(|h_j|^2\right)'\left(\frac{1}{\theta}\right)'' + 2|h_j|^2\left(\frac{1}{\theta}\right)''' \\ = 0.\end{aligned} \tag{51}$$

Equation (49) is equivalent to (50) and (51). Equation (50) is also

$$2|h_j|^2A\left(\frac{1}{\theta}\right) + 2i[h_j,\bar{h}_j]B\left(\frac{1}{\theta}\right) + 3i\left([h_j,\bar{h}_j]\left(\frac{1}{\theta}\right)'\right)' = 0. \tag{52}$$

The first two terms in (52) are exact derivatives. Therefore, we set

$$A = ia\left([h_j,\bar{h}_j]\left(\frac{1}{\theta}\right)\right)', \quad B = a\left(|h_j|^2\left(\frac{1}{\theta}\right)\right)', \tag{53}$$

where $a = a(t)$ is a real valued function and does not depend on x.

Then (52) is equivalent to

$$2a\left([h_j,\bar{h}_j]|h_j|^2\left(\frac{1}{\theta^2}\right)\right)' + 3\left([h_j,\bar{h}_j]\left(\frac{1}{\theta}\right)'\right)' = 0. \tag{54}$$

Since $[h_j,\bar{h}_j] = 0$ at $x = 0$, (54) gives $\theta' = \frac{2}{3}a|h_j|^2$, then

$$\theta(x,t) = 1 + \frac{2}{3}a(t)\int_0^x |h_j(s)|^2\,ds. \tag{55}$$

Note that any other choice of $c(t)$ instead of 1 in (55) would affect only $a(t)$ and $b(t)$. Next we determine $a(t)$ in terms of $b(t)$ using (48) and (55). From (53)

$$\frac{d}{dt}\begin{pmatrix} A \\ B \end{pmatrix} = \frac{d}{dx}\left(\frac{1}{\theta}\frac{d}{dt}a - \frac{a}{\theta^2}\frac{d}{dt}\theta\right)\begin{pmatrix} i[h_j,\bar{h}_j] \\ |h_j|^2 \end{pmatrix}.$$

Then from (48) we get

$$\frac{1}{\theta}\frac{d}{dt}a - \frac{a}{\theta^2}\frac{d}{dt}\theta = \frac{3}{2}b^2. \tag{56}$$

Using (55) and (56) we obtain

$$\frac{d}{dt}a = \frac{3b^2}{2\theta^2}. \tag{57}$$

Besides it is easy to check that

$$\left|\frac{h_j}{\theta}\right|^2 = |h_j|^2 - \frac{2}{3}a\left(\frac{1}{\theta}\left(\int_0^x |h_j|^2\right)^2\right)' ds. \tag{58}$$

Then (58) gives

$$\left\|\frac{h_j}{\theta}\right\|^2_{L^2_{\mathbf{R}}[0,1]} = \frac{1}{1+\frac{2}{3}a}. \tag{59}$$

Therefore, (46) and (59) give

$$1 + \frac{2}{3}a = b^2. \tag{60}$$

Following (59), $a(0) = 0$. Combining (57) and (60) and looking for $a(t)$ with $a(0) = 0$ we get

$$\frac{2}{3}a(t) = e^t - 1, \quad b(t) = e^{\frac{t}{2}}. \tag{61}$$

Therefore, (50) is satisfied with

$$\theta(x,t) = 1 + (e^t - 1)\int_0^x |h_j(s)|^2\, ds, \tag{62}$$

$$A(x,t) = \frac{3}{2}i(e^t - 1)\frac{d}{dx}\frac{[h_j(x), \bar{h}_j(x)]}{\theta(x,t)}, \tag{63}$$

$$\begin{aligned} B(x,t) &= \frac{3}{2}(e^t - 1)\frac{d}{dx}\frac{|h_j(x)|^2}{\theta(x,t)} \\ &= \frac{3}{2}\frac{d^2}{dx^2}\log\theta(x,t). \end{aligned} \tag{64}$$

The function $\theta(x,t) > 0\ \forall\,(x,t) \in [0,1] \times \mathbf{R}$. Therefore, (50) is verified with (62), (63), and (64). It remains to check that (51) is satisfied with (62) and (64). From (53) and (55)

$$B = \frac{3}{2}\left(\frac{\theta'}{\theta}\right)'. \tag{65}$$

Using (65) and $|h_j|^2 = \frac{3}{2a}\theta'$, the left-hand side of (51) is, after expressing $\left(\frac{1}{\theta}\right)'''$, $\left(\frac{1}{\theta}\right)''$, $\left(\frac{1}{\theta}\right)'$,

$$\begin{aligned}
&\frac{9}{2a}\frac{\theta'}{\theta}\left(\frac{\theta'''}{\theta} - 3\frac{\theta''\theta'}{\theta^2} + 2\left(\frac{\theta'}{\theta}\right)^3\right) + \frac{9}{2a}\frac{\theta''}{\theta}\left(\frac{\theta''}{\theta} - \left(\frac{\theta'}{\theta}\right)^2\right)\\
&- \frac{9}{a}\left(\frac{\theta'}{\theta}\right)^2\left(\frac{\theta''}{\theta} - \left(\frac{\theta'}{\theta}\right)^2\right) - \frac{6}{a}q_0\left(\frac{\theta'}{\theta}\right)^2 - 3\frac{\theta'}{\theta^2}(\bar{h}_j''h_j + \bar{h}_j h_j'') \qquad (66)\\
&+ \frac{9}{2a}\theta''\left(2\frac{(\theta')^2}{\theta^3} - \frac{\theta''}{\theta^2}\right) + \frac{3}{a}\theta'\left(-6\frac{(\theta')^3}{\theta^4} + 6\frac{\theta'\theta''}{\theta^3} - \frac{\theta'''}{\theta^2}\right).
\end{aligned}$$

In (66) the terms $\frac{(\theta')^2\theta''}{\theta^3}$, $(\frac{\theta'}{\theta})^4$, $(\frac{\theta''}{\theta})^2$ vanish and the remaining terms are

$$\frac{3}{2a}\frac{\theta'''}{\theta^2}\theta' - \frac{6q_0}{a}\left(\frac{\theta'}{\theta}\right)^2 - 3\frac{\theta'}{\theta^2}(\bar{h}_j''h_j + \bar{h}_j h_j''). \qquad (67)$$

Using

$$\theta''' = \frac{2}{3}a(\bar{h}_j''h_j + 2\bar{h}_j'h_j' + \bar{h}_j h_j''),$$

(67) is equal to

$$\begin{aligned}
&- 2\frac{\theta'}{\theta^2}(\bar{h}_j''h_j - |h_j'|^2 + \bar{h}_j h_j'' + 2q_0|h_j|^2)\\
&= 2i\frac{\theta'}{\theta^2}w(\bar{h}_j, h_j). \qquad (68)
\end{aligned}$$

Following (39) the derivative of $w(\bar{h}_j, h_j)$ with respect to x vanishes. Then the assumption $h_j'(0) = 0$ and (2) show that (68) is zero for every $x \in [0,1]$. Then (51) is verified. □

REFERENCES

[A1] L. Amour, *Inverse spectral theory for the AKNS system with separated boundary conditions*, Inverse Problems, 9 (1993), pp. 507–523.

[A2] L. Amour, *Determination of a third-order operator from two of its spectra*, SIAM J. Math. Anal., 30 (1999), pp. 1010–1028.

[AG] L. Amour and J.-C. Guillot, *Isospectral sets for AKNS systems with generalized periodic boundary conditions*, Geom. Funct. Anal., 6 (1996), pp. 1–27.

[CM] C. F. Coleman and J. R. McLaughlin, *Solution of the inverse spectral problem for an impedance with integrable derivative. Parts* I, II, Comm. Pure Appl. Math., 46 (1993), pp. 145–212.

[FIT] A. Finkel, E. Isaacson, and E. Trubowitz, *An explicit solution of the inverse periodic problem for Hill's equation*, SIAM J. Math. Anal., 18 (1987), pp. 46–53.

[GST] F. Gesztezy, B. Simon, and G. Teschl, *Spectral deformations of one-dimensional Schrödinger operators*, J. Anal. Math., 70 (1996), pp. 267–324.

[GR] J.-C. Guillot and J. V. Ralston, *Inverse spectral theory for a singular Sturm-Liouville operator on* [0,1], J. Differential Equations, 76 (1988), pp. 353–373.

[IT] E. L. Isaacson and E. Trubowitz, *The inverse Sturm-Liouville problem* I, Comm. Pure Appl. Math., 36 (1983), pp. 767–783.

[M] H. P. McKean, *Boussinesq's equation on the circle*, Comm. Pure Appl. Math., 34 (1981), pp. 599–691.

[MT] H. P. McKean and E. Trubowitz, *The spectral class of the quantum-mechanical harmonic oscillator*, Comm. Math. Phys., 82 (1982), pp. 471–495.

[PT] J. Pöschel and E. Trubowitz, *Inverse Spectral Theory*, Academic Press, Boston, MA, 1987.

[RT] J. Ralston and E. Trubowitz, *Isospectral sets for boundary value problem on the unit interval*, Ergodic Theory Dynam. Systems, 8 (1988), pp. 301–358.

SIAM J. MATH. ANAL.
Vol. 32, No. 6, pp. 1390–1403

GLOBAL EXISTENCE OF SMALL SOLUTIONS TO THE QUADRATIC NONLINEAR SCHRÖDINGER EQUATIONS IN TWO SPACE DIMENSIONS*

NAKAO HAYASHI† AND PAVEL I. NAUMKIN‡

Abstract. We study a global existence in time of small solutions to the quadratic nonlinear Schrödinger equation in two space dimensions,

$$\begin{cases} i\partial_t u + \frac{1}{2}\Delta u = \mathcal{N}(u), & (t,x) \in \mathbf{R}\times\mathbf{R}^2, \\ u(0,x) = u_0(x), & x \in \mathbf{R}^2, \end{cases} \tag{0.1}$$

where

$$\mathcal{N}(u) = \sum_{j,k=1}^{2} \left(\lambda_{jk}(\partial_{x_j} u)(\partial_{x_k} u) + \mu_{jk}(\partial_{x_j}\bar{u})(\partial_{x_k}\bar{u})\right),$$

$\lambda_{jk}, \mu_{jk} \in \mathbf{C}$. We prove that if the initial data u_0 satisfy some analyticity and smallness conditions in a suitable norm, then the solution of the Cauchy problem (0.1) exists globally in time. Furthermore we prove the existence of the usual scattering states.

Key words. nonlinear Schrödinger equations, global existence, quadratic nonlinearities, two spatial dimensions

AMS subject classification. 35Q35

PII. S0036141000372532

1. Introduction.

In this paper we prove the global existence of small analytic solutions and the existence of the usual scattering states to the Cauchy problem for the derivative nonlinear Schrödinger equation

$$\begin{cases} i\partial_t u + \frac{1}{2}\Delta u = \mathcal{N}(u), & (t,x) \in \mathbf{R}\times\mathbf{R}^n, \\ u(0,x) = u_0(x), & x \in \mathbf{R}^n, \end{cases} \tag{1.1}$$

with quadratic nonlinearity

$$\mathcal{N}(u) = \sum_{j,k=1}^{n} \left(\lambda_{jk}\left(\partial_{x_j} u\right)\left(\partial_{x_k} u\right) + \mu_{jk}\left(\partial_{x_j}\bar{u}\right)\left(\partial_{x_k}\bar{u}\right)\right),$$

where $\lambda_{jk}, \mu_{jk} \in \mathbf{C}$ when the space dimension $n = 2$. Global existence of small solutions to nonlinear Schrödinger equations with quadratic nonlinearities was first studied in papers [12], [13], [15] under the condition

$$\frac{\partial \mathcal{N}}{\partial(\partial_{x_j} u)} \quad \text{is pure imaginary} \tag{1.2}$$

and the space dimension $n \geq 5$. Condition (1.2) was removed in [8], where the global existence of small solutions to the Cauchy problem (1.1) with general nonlinearity

*Received by the editors May 22, 2000; accepted for publication (in revised form) November 27, 2000; published electronically April 6, 2001.
http://www.siam.org/journals/sima/32-6/37253.html

†Department of Applied Mathematics, Science University of Tokyo, Tokyo 162-8601, Japan (nhayashi@rs.kagu.sut.ac.jp). Current address: Department of Mathematics, Graduate School of Science, Osaka University, Toyonaka, Osaka 560-0043, Japan.

‡Instituto de Física y Matemáticas, Universidad Michoacana, AP 2-82, CP 58040, Morelia, Michoacán, México (pavelni@zeus.ccu.umich.mx).

was proved for any space dimension $n \geq 5$. Note that the problem on the global existence of solutions becomes more difficult in the case of low space dimensions. In [11] a global existence theorem with the condition

$$\left|\frac{\partial \mathcal{N}}{\partial u}\right| + \left|\frac{\partial \mathcal{N}}{\partial \bar{u}}\right| \leq C\,|\nabla u| \tag{1.3}$$

was obtained for the case $n = 3, 4$. In [7] the problem was solved for the case $n = 4$ with condition (1.2). Recently, the nonlinear term

$$\mathcal{N} = \lambda u^2 + \mu \bar{u}^2, \tag{1.4}$$

where $\lambda, \mu \in \mathbf{C}$, was considered in [10] and a global existence theorem for small solutions was shown in three space dimensions. However, for low spatial dimensions there are only a few results. In the exceptional case

$$\mathcal{N} = (\nabla u)^2, \tag{1.5}$$

(1.1) can be linearized by the Hopf–Cole transformation and the solution can be written explicitly, so the global existence and the asymptotic properties of the solution can be studied; see paper [14]. If we can prove that the existence time T of solutions has a representation $T = O\left(\exp\left(\frac{C}{\varepsilon}\right)\right)$, where $\varepsilon > 0$ is the size of the initial data in a suitable norm, then we call the result an almost global existence of solutions. An almost global existence theorem was proved in [5] for the case $n = 3$. In the case of two space dimensions the almost global existence of small analytic solutions to the Cauchy problem (1.1) with the nonlinearity

$$\mathcal{N}(u) = \sum_{j,k=1}^{2} \left(\lambda_{jk} \left(\partial_{x_j} u\right) \left(\partial_{x_k} u\right) + \mu_{jk} \left(\partial_{x_j} \bar{u}\right) \left(\partial_{x_k} \bar{u}\right) + \nu_{jk} \left(\partial_{x_j} u\right) \left(\partial_{x_k} \bar{u}\right)\right)$$

was studied in paper [9]. Moreover, it was proved that the global existence of small analytic solutions to the problem (1.1) in the case of the nonlinearity is

$$\mathcal{N} = \lambda \left(\partial_{x_1} u \partial_{x_2} \bar{u} - \partial_{x_1} \bar{u} \partial_{x_2} u\right), \tag{1.6}$$

where $\lambda \in \mathbf{C}$. In order to state the result from [4] we need the condition

$$\mathcal{N}(u) = \sum_{j,k=1}^{2} \mu_{jk} \left(\partial_{x_j} \bar{u}\right) \left(\partial_{x_k} \bar{u}\right). \tag{1.7}$$

In [4] the global existence of small solutions in the usual Sobolev spaces was established for the case (1.7) by using the method of normal forms introduced in [16]. We note that the method used in [4] does not work for the case $\lambda_{jk} \neq 0$. For the reader's convenience we use Tables 1.1 and 1.2 to show the previous works on the global solvability of nonlinear Schrödinger equations with quadratic nonlinearities. To make these tables we introduce the following conditions:

$$\mathcal{N} \quad \text{does not depend on} \quad \nabla u \quad \text{and} \quad \nabla \bar{u}. \tag{1.8}$$

Note that the nonlinearity (1.4) is a particular case of (1.8). Also, we define the condition

$$\mathcal{N} \quad \text{does not depend on} \quad u \quad \text{and} \quad \bar{u}. \tag{1.9}$$

TABLE 1.1
Global existence for quadratic nonlinear Schrödinger equations.

$n \diagdown \mathcal{N}$	General	(1.2)	(1.3)	(1.2) and (1.3)	(1.8)	(1.9)
3	–	–	[11]	[6]	–	[8],[9],[11]
4	–	[7]	[11]	[6]	[17]	
$n \geq 5$	[8],[9],[11]			[12],[13],[15]		

$n \diagdown \mathcal{N}$	(1.4)	(1.5)	(1.6)	(1.7)
1	–	[14]	–	–
2	–		[9]	[4]
3	[10]			

TABLE 1.2
Almost global existence for quadratic nonlinear Schrödinger equations.

$n \diagdown \mathcal{N}$	(1.2)	(1.8)	(1.2) and (1.3)	(1.9)
1	–	–	–	–
2	–	–	[7]	[9]
3	[5]	[17]		

As the consequences of (1.9) we have nonlinearity (1.6) and the Hopf–Cole case (1.5).

As we see from Tables 1.1 and 1.2 there are no global existence and time decay results in the one-dimensional case (except the Hopf–Cole case (1.5)). The estimate $T = O\left(\varepsilon^{-6}\right)$ for time existence of solutions was shown in [3] in one spatial dimension, if the nonlinearity $\mathcal{N} = i(\bar{u}_x)^2$ by using the method of normal forms.

The difficulty in the study of the large time asymptotic behavior of solutions to the Cauchy problem (1.1) is that the quadratic nonlinearity in two space dimensions is critical; that is, it decays in time with the same speed as the linear terms in the equation and cannot be omitted in the first approximation of the perturbation theory. The special oscillating structure of the nonlinearity must be taken into account to show that the solution has the usual scattering properties. Another difficulty is that the structure of the quadratic nonlinearity under consideration is not self-conjugate; that is, it does not satisfy the property $\mathcal{N}\left(ue^{i\theta}\right) = e^{i\theta}\mathcal{N}(u)$ for all $\theta \in \mathbf{R}$. This fact does not allow us to estimate the operator $\mathcal{J} = x + it\nabla$ via the standard energy type methods. To overcome this obstacle and also the so-called difficulty of the derivatives loss, we apply the analytic functional spaces. The important fact is that the nonlinearity under consideration has the type (1.9). This helps us to evaluate large time asymptotics of the nonlinearity, since the remainder terms can be estimated by the operator $\mathcal{Q} = x \cdot \nabla + it\Delta = \mathcal{J} \cdot \nabla$ through the operator $\mathcal{P} = x \cdot \nabla + 2t\partial_t$.

To state our result precisely, we now give *notation and function spaces.* We denote $\partial_{x_j} = \frac{\partial}{\partial x_j}$ and $\partial^\alpha = \partial_{x_1}^{\alpha_1}\partial_{x_2}^{\alpha_2}$, where $\alpha \in (\mathbf{N} \cup \{0\})^2$. We define the differential operators $\mathcal{P} = x \cdot \nabla + 2t\partial_t$, $\mathcal{Q} = x \cdot \nabla + it\Delta$, the vector $\Omega = \left(\Omega^{(j,k)}\right)_{(j,k=1,2)}$, where the operators $\Omega^{(j,k)} = x_j\partial_k - x_k\partial_j$ act as the angular derivatives, and the vector $\mathcal{J} = (\mathcal{J}_j)_{(j=1,2)}$ with components $\mathcal{J}_j = x_j + it\partial_j$. These operators help us to obtain the time decay properties of the linear Schrödinger evolution group $\mathcal{U}(t)\phi = \frac{1}{2\pi it}\int e^{\frac{i}{2t}(x-y)^2}\phi(y)\,dy = \mathcal{F}^{-1}e^{-\frac{it}{2}\xi^2}\mathcal{F}\phi$, where $\mathcal{F}\phi \equiv \hat{\phi}(\xi) = \frac{1}{2\pi}\int e^{-i(x\cdot\xi)}\phi(x)\,dx$ denotes the Fourier transform of the function $\phi(x)$ and $\mathcal{F}^{-1}$ is the inverse Fourier transformation defined by $\mathcal{F}^{-1}\phi \equiv \check{\phi}(x) = \frac{1}{2\pi}\int e^{i(x\cdot\xi)}\phi(\xi)\,d\xi$. We also denote the extended vectors $\Gamma = (\mathcal{P}, \Omega, \nabla)$, $\widetilde{\Gamma} = (\mathcal{P} + 2, \Omega, \nabla)$, and $\Theta = (\mathcal{Q}, \Omega, \nabla)$. We have the fol-

lowing relations: $\mathcal{Q}=\mathcal{P}-2it\mathcal{L}=\mathcal{J}\cdot\nabla=\mathcal{U}(t)\,x\mathcal{U}(-t)\cdot\nabla=it\mathcal{M}(t)\,\nabla\overline{\mathcal{M}(t)}\cdot\nabla$, where $\mathcal{M}(t)=e^{ix^2/2t}$, $\mathcal{L}=i\partial_t+\frac{1}{2}\Delta$. We use freely in the paper the following commutation relations: $[\mathcal{Q},\nabla]=[\mathcal{P},\nabla]=-\nabla$, $[\mathcal{Q},\mathcal{J}]=[\mathcal{P},\mathcal{J}]=\mathcal{J}$, $[\mathcal{P},\mathcal{Q}]=[\Omega,\mathcal{P}]=[\Omega,\mathcal{Q}]=0$, $[\partial_k,\mathcal{J}_l]=\delta_l^{(k)}$, $[\Omega_x^{(j,k)},\partial_l]=\delta_l^{(k)}\partial_j-\delta_l^{(j)}\partial_k$, where $\delta_j^{(k)}=1$ if $j=k$ and $\delta_j^{(k)}=0$ if $j\neq k$. We note that the free Schrödinger evolution group $\mathcal{U}(t)$ can be represented in the following manner: $\mathcal{U}(t)=\mathcal{M}(t)\,\mathcal{D}(t)\,\mathcal{F}M(t)$, where $\mathcal{M}(t)=\exp(ix^2/2t)$; the dilation operator is $(\mathcal{D}(t)\,\varphi)(x)=\frac{i}{t}\varphi\left(\frac{x}{t}\right)$. The inverse free Schrödinger evolution group is written as $\mathcal{U}(-t)=-\mathcal{M}(-t)\,i\mathcal{F}^{-1}\mathcal{D}\left(\frac{1}{t}\right)\mathcal{M}(-t)$, where $\mathcal{D}^{-1}(t)=-i\mathcal{D}\left(\frac{1}{t}\right)$ is the inverse dilation operator. We denote the usual Lebesgue space by $\mathbf{L}^p\left(\mathbf{R}^2\right)$ with the norm $\|\phi\|_p=\left(\int_{\mathbf{R}^2}|\phi(x)|^p\,dx\right)^{1/p}$ if $1\leq p<\infty$ and $\|\phi\|_\infty=\operatorname{ess.sup}\left\{|\phi(x)|\,;x\in\mathbf{R}^2\right\}$ if $p=\infty$. For simplicity we write $\|\cdot\|=\|\cdot\|_2$. The weighted Sobolev space is defined by $\mathbf{H}_p^{m,k}\left(\mathbf{R}^2\right)=\{\phi\in\mathbf{L}^2\left(\mathbf{R}^2\right):\|\langle x\rangle^k\langle i\nabla\rangle^m\phi\|_p<\infty\}$, where $m,k\in\mathbf{R}^+$, $1\leq p\leq\infty$, $\langle x\rangle=\sqrt{1+x^2}$ are the Japanese brackets. We also denote $\mathbf{H}^{m,k}\left(\mathbf{R}^2\right)=\mathbf{H}_2^{m,k}\left(\mathbf{R}^2\right)$ and the norm $\|\phi\|_{m,k}=\|\phi\|_{m,k,2}$. We now define the analytic function space

$$\mathbf{A}^{m,p}(t)=\left\{\phi\in\mathbf{L}^p\left(\mathbf{R}^2\right);\|\phi\|_{\mathbf{A}^{m,p}}=\sum_{|\beta|\leq m}\sum_{\alpha}\frac{b^{|\alpha|}}{\alpha!}\left\|\Gamma^{\alpha+\beta}\phi\right\|_p<\infty\right\},$$

where $\Gamma=\Gamma(t)=(\mathcal{P},\Omega,\nabla)$, $b=b(t)=b_\infty+(b_0-b_\infty)\left(\log(e+t)\right)^{-\gamma}$, $0<b_\infty<b_0<1$, and $\gamma>0$ is sufficiently small. Similarly we write

$$\widetilde{\mathbf{A}}^{m,p}(t)=\left\{\phi\in\mathbf{L}^p\left(\mathbf{R}^2\right);\|\phi\|_{\widetilde{\mathbf{A}}^{m,p}}=\sum_{|\beta|\leq m}\sum_{\alpha}\frac{b^{|\alpha|}}{\alpha!}\left\|\widetilde{\Gamma}^{\alpha+\beta}\phi\right\|_p<\infty\right\}.$$

Here the summation is over all admissible multi-indices α. We often use the summations convention if it does not cause confusion. For simplicity we write $\mathbf{A}^m(t)=\mathbf{A}^{m,2}(t)$ and $\widetilde{\mathbf{A}}^m(t)=\widetilde{\mathbf{A}}^{m,2}(t)$. By $[s]$ we denote the largest integer less than or equal to s. Let $\mathbf{C}(\mathbf{I};\mathbf{B})$ be the space of continuous functions from a time interval $\mathbf{I}$ to a Banach space $\mathbf{B}$. Different positive constants we denote by the same letter C. Our basic estimates of the solution are in the functional space $\mathbf{X}=\left\{u\in\mathbf{C}\left(\mathbf{R};\mathbf{L}^2\right);\|u\|_{\mathbf{X}}<\infty\right\}$, where

$$\begin{aligned}\|u\|_{\mathbf{X}}&=\sup_{t>0}\|u(t)\|_{\mathbf{A}^3(t)}+\sup_{t>0}t^{-1-\eta}\sum_{|\gamma|\leq1}\|\mathcal{J}^\gamma u(t)\|_{\mathbf{A}^2(t)}\\&+\sum_{|\gamma|=1}\int_0^\infty\|\Theta^\gamma u\|_{\mathbf{A}^3(t)}\,|b'|\,dt+\sum_{|\gamma|=1,|\sigma|\leq1}\int_1^\infty\|\Theta^\gamma\mathcal{J}^\sigma u\|_{\mathbf{A}^3(t)}\,\frac{|b'|\,dt}{t^{1+\eta}}\\&+\sup_{t>0}t^{1-2\eta}\sum_{\alpha}\frac{b^{|\alpha|}}{\alpha!}\|\partial_t\mathcal{V}(-t)\,\mathcal{K}\Gamma^\alpha\nabla u(t)\|_\infty\\&+\sum_{|\delta|\leq3}\int_1^\infty\sum_{\alpha}\frac{b^{|\alpha|}}{\alpha!}\left\|\partial_t\mathcal{V}(-t)\,\mathcal{K}\Gamma^{\alpha+\delta}u(t)\right\|t^{2\eta-\frac{1}{2}}dt;\end{aligned}$$

here $\mathcal{K}=\mathcal{F}\mathcal{M}\mathcal{U}(-t)$ and $\eta>0$ is sufficiently small. Now we state our result.

THEOREM 1.1. *We assume that the initial data $u_0\in\mathbf{A}^3(0)$ are such that $x_ju_0\in\mathbf{A}^2(0)$ for $j=1,2$ and the norm $\|u_0\|_{\mathbf{A}^3(0)}+\|x_1u_0\|_{\mathbf{A}^2(0)}+\|x_2u_0\|_{\mathbf{A}^2(0)}=\varepsilon$ is sufficiently small. Then there exists a unique global solution of the Cauchy problem*

(1.1) *such that* $u(t,x) \in \mathbf{A}^3(t)$ *for all* $t \in \mathbf{R}$. *Moreover, there exist unique functions* $u^{\pm} \in \mathbf{L}^2(\mathbf{R}^2)$ *such that*

$$\left\| \mathcal{U}(-t)\, u(t) - u^{\pm} \right\| \le C\, |t|^{-\omega},$$

as $t \to \pm\infty$, *where* $0 < \omega < \frac{1}{2}$.

We assumed in the theorem that $0 < b_\infty < b_0 < 1$. This ensures that the function space $\mathbf{A}^3(0)$ for the initial data is not empty: as in [1], [2], we can see that our result is valid for the initial function ϕ, which has an analytic continuation Φ to the domain

$$\begin{aligned} \Pi = \big\{ z \in \mathbf{C}^2;\ & z_j = x_j + i y_j,\ x_j \in \mathbf{R}, \\ & -a - |x_j| \tan\vartheta < y_j < a + |x_j| \tan\vartheta,\ j = 1,2 \big\} \end{aligned}$$

such that

$$\int\int_{\Pi} |\Phi(z)|^2\, dxdy < \infty,$$

where $\vartheta \in \left(0, \frac{\pi}{2}\right)$, $\sin\vartheta = c$, and $a, c \in (b_0, 1)$. For example, we can take $1/\left(1+x^4\right)$, e^{-x^2} as the initial data for the Cauchy problem (1.1).

We conclude this section with the following remark. The method of the present paper can also be applied for the proof of the global existence of small solutions to the system of derivative nonlinear Schrödinger equations

$$\begin{cases} i\partial_t v + \frac{1}{2}\Delta v = \partial_{x_1} \mathcal{N}(v,w), & (t,x) \in \mathbf{R} \times \mathbf{R}^2, \\ i\partial_t w + \frac{1}{2}\Delta w = \partial_{x_2} \mathcal{N}(v,w), & (t,x) \in \mathbf{R} \times \mathbf{R}^2, \\ v(0,x) = v_0(x), w(0,x) = w_0(x) & x \in \mathbf{R}^2, \end{cases}$$

where

$$\mathcal{N}(v,w) = \lambda_1 v^2 + \lambda_2 \overline{v}^2 + \lambda_3 w^2 + \lambda_4 \overline{w}^2 + \lambda_5 vw + \lambda_6 \overline{vw},$$

where $\lambda_j \in \mathbf{C}$, $j = 1, \dots, 6$. This system can be obtained from (1.1) via differentiating with respect to x_1 and x_2 and putting $v = \partial_{x_1} u$ and $w = \partial_{x_2} u$.

The rest of the paper we organize as follows. In section 2 we give some preliminary estimates in the analytic functional spaces $\mathbf{A}^{m,p}$. Section 3 is devoted to the proof of Theorem 1.1.

2. Preliminary estimates.

LEMMA 2.1. *Let* $\phi \in \mathbf{A}^{m,p}(t)$. *Then we have*

$$\|\phi\|_{\tilde{\mathbf{A}}^{m,p}(t)} \le e^{2b} \|\phi\|_{\mathbf{A}^{m,p}(t)}.$$

Proof. We have

$$\begin{aligned} \sum_{l=0}^{\infty} \frac{b^l}{l!} \left\| (\mathcal{P}+2)^l \phi \right\|_p &\le \sum_{l=0}^{\infty} \sum_{k=0}^{l} \frac{(2b)^k}{k!} \frac{b^{l-k}}{(l-k)!} \left\| \mathcal{P}^{l-k} \phi \right\|_p \\ &\le e^{2b} \sum_{l=0}^{\infty} \frac{b^l}{l!} \left\| \mathcal{P}^l \phi \right\|_p, \end{aligned}$$

whence the result of the lemma follows. □

Denote by $C_m^k = \frac{m!}{(m-k)!k!}$ the binomial coefficients.

LEMMA 2.2. *We have the following commutation relations:*

$$\left[\partial_{x_j}^l, \mathcal{J}_{x_j}\right] = l\partial_{x_j}^{l-1}, \quad \left[\mathcal{J}_j^l, \partial_{x_j}\right] = -l\mathcal{J}_{x_j}^{l-1},$$

$$\mathcal{P}^l \mathcal{J}_{x_j} = \sum_{0\le m\le l} C_l^m \mathcal{J}_{x_j} \mathcal{P}^{l-m}, \quad \mathcal{P}^l \partial_{x_j} = \sum_{0\le m\le l} C_m^l \left(-1\right)^m \partial_{x_j} \mathcal{P}^{l-m}, \tag{2.1}$$

$$\mathcal{J}_{x_j} \mathcal{P}^l = \sum_{0\le m\le l} C_l^m \left(-1\right)^m \mathcal{P}^{l-m} \mathcal{J}_{x_j}, \quad \partial_{x_j} \mathcal{P}^l = \sum_{0\le m\le l} C_l^m \mathcal{P}^{l-m} \partial_{x_j}, \tag{2.2}$$

$$\begin{aligned} \Omega_{x_j x_k}^l \partial_{x_j} = &\sum_{0\le 2m\le l} \left(-1\right)^m C_l^{2m} \partial_{x_j} \Omega_{x_j x_k}^{l-2m} \\ &+ \sum_{0\le 2m+1\le l} \left(-1\right)^{m+1} C_l^{2m+1} \partial_{x_k} \Omega_{x_j x_k}^{l-2m-1}, \end{aligned} \tag{2.3}$$

$$\begin{aligned} \partial_{x_j} \Omega_{x_j x_k}^l = &\sum_{0\le 2m\le l} \left(-1\right)^{m+1} C_l^{2m} \Omega_{x_j x_k}^{l-2m} \partial_{x_j} \\ &+ \sum_{0\le 2m+1\le l} \left(-1\right)^m C_l^{2m+1} \Omega_{x_j x_k}^{l-2m-1} \partial_{x_k}. \end{aligned} \tag{2.4}$$

Formulas (2.1) *and* (2.2) *also hold if we replace* $\mathcal{P}$ *by* $\mathcal{Q}$*, and formulas* (2.3) *and* (2.4) *are valid if we replace* ∂_{x_j} *by* $\mathcal{J}_{x_j}$*.*

For the proof, see [9, Lemma 2.6].

The following lemma is an analytic version of Lemma 2.2 from paper [6]. It will be used in many stages of the proof of the result.

LEMMA 2.3. *We have the estimate*

$$\left\|\mathcal{J}\nabla\phi\right\|_{\mathbf{A}^{m,p}(t)} \le C \sum_{|\alpha|=1} \left\|\Theta^\alpha \phi\right\|_{\mathbf{A}^{m,p}(t)}.$$

For the proof, see [9, Lemma 2.5].

The next lemma is obtained via the application of the commutation relations of Lemma 2.2. (See [9, Lemma 2.7].)

LEMMA 2.4. *The inequalities*

$$C_1 \left\|\partial_{x_j}\phi\right\|_{\mathbf{A}^{m,p}(t)} \le \sum_{|\beta|\le m} \sum_\alpha \frac{b^{|\alpha|}}{\alpha!} \left\|\partial_{x_j} \Gamma^{\alpha+\beta}\phi\right\|_p \le C_2 \left\|\partial_{x_j}\phi\right\|_{\mathbf{A}^{m,p}(t)}$$

and

$$\begin{aligned} C_1 \left\|\mathcal{J}_{x_j}\phi\right\|_{\mathbf{A}^{m,p}(t)} &\le \sum_{|\beta|\le m} \sum \frac{b^{|\alpha|}}{\alpha!} \left\|\mathcal{J}_{x_j} \Gamma^{\alpha+\beta}\phi\right\|_p + \left\|\phi\right\|_{\mathbf{A}^{m,p}(t)} \\ &\le C_2 \left(\left\|\mathcal{J}_{x_j}\phi\right\|_{\mathbf{A}^{m,p}(t)} + \left\|\phi\right\|_{\mathbf{A}^{m,p}(t)}\right) \end{aligned}$$

are true for all $t > 0$*, where* C_1*,* $C_2 > 0$*.*

To estimate the nonlinearity in the analytic functional spaces we use the following result. (See [9, Lemma 2.8] for the proof.)

LEMMA 2.5. *We have the estimate*

$$\|\phi\psi\|_{\widetilde{\mathbf{A}}^{0,p}(t)} \le C \|\phi\|_{\mathbf{A}^{0,q}(t)} \|\psi\|_{\mathbf{A}^{0,r}(t)}$$

for all $t > 0$, *where* $\frac{1}{p} = \frac{1}{q} + \frac{1}{r}$.

Let $1 \le q, r \le \infty$, $j, m \in \mathbf{N} \cup \{0\}$ *be such that* $0 \le j \le m$. *Then the Sobolev embedding inequality is true in the analytic functional spaces*

$$\sum_{|\alpha|=j} \|\partial^\alpha \phi\|_{\mathbf{A}^{0,p}(t)} \le C \sum_{|\beta|=m} \left\|\partial^\beta \phi\right\|^a_{\mathbf{A}^{0,r}(t)} \|\phi\|^{1-a}_{\mathbf{A}^{0,q}(t)}$$

for all $t > 0$, *where* $\frac{1}{p} = \frac{j}{2} + a\left(\frac{1}{r} - \frac{m}{2}\right) + \frac{1-a}{q}$ *and* a *is such that* $\frac{j}{m} \le a < 1$ *if* $m - j - \frac{2}{r} \in \mathbf{N} \cup \{0\}$, *and* $\frac{j}{m} \le a \le 1$ *otherwise.*

We define the evolution operator

$$\mathcal{V}(t)\varphi = \mathcal{F}^{-1} e^{\frac{i\xi^2}{2t}} \mathcal{F}\varphi = \frac{t}{2i\pi} \int e^{-i\frac{t}{2}(\xi - y)^2} \varphi(y)\, dy$$

and $\mathcal{K} = \mathcal{F}\mathcal{M}\mathcal{U}(-t)$. By a direct calculation we see that

$$\mathcal{V}(-t)\left(E^{\omega-1}\varphi\right) = \mathcal{D}_\omega E^{\omega(\omega-1)} \mathcal{V}(-\omega t)\varphi \tag{2.5}$$

with $\mathcal{D}_\omega \varphi = \frac{1}{\omega}\varphi\left(\frac{\xi}{\omega}\right)$ and $E = e^{it\xi^2/2}$, where $\omega \ne 0$. We need the following lemma to get the decay estimates of the solution for large time.

LEMMA 2.6. *The estimates*

$$\left\|\mathcal{D}_\omega E^{\omega(\omega-1)} (\mathcal{V}(-\omega t) - 1)(\mathcal{K}\phi)(\mathcal{K}\psi)\right\| + \left\|\mathcal{D}_\omega E^{\omega(\omega-1)} (\mathcal{K}\psi)(\mathcal{V}(-\omega t) - 1)(\mathcal{K}\phi)\right\| \le C t^{\eta - \frac{1}{2}} \sum_{|\alpha|\le 1, |\beta| \le 1} \|\mathcal{J}^\alpha \phi\| \left\|\mathcal{J}^\beta \psi\right\|$$

and

$$\left\|\mathcal{D}_\omega E^{\omega(\omega-1)} (\mathcal{V}(-\omega t) - 1)(\mathcal{K}\phi)(\mathcal{K}\psi)\right\|_\infty + \left\|\mathcal{D}_\omega E^{\omega(\omega-1)} (\mathcal{K}\psi)(\mathcal{V}(-\omega t) - 1)(\mathcal{K}\phi)\right\|_\infty \le C \left(\sum_{|\alpha|\le 1, |\beta|\le 1} \|\mathcal{J}^\alpha \phi\|^{1-\eta} \left\|\mathcal{J}^\beta \psi\right\|^{1-\eta}\right) \sum_{|\alpha|\le 2, |\beta|\le 2} \|\mathcal{J}^\alpha \phi\|^\eta \left\|\mathcal{J}^\beta \psi\right\|^\eta$$

are valid for all $t > 0$, *where* $\omega \ne 0$, $\eta > 0$ *is sufficiently small.*

Proof. Note that $\|\mathcal{K}\phi\|_{1,0} = \|\mathcal{U}(-t)\phi\|_{0,1} \le C \sum_{|\alpha|\le 1} \|\mathcal{J}^\alpha \phi\|$. By the Sobolev embedding theorem we obtain

$$\left\|\mathcal{D}_\omega E^{\omega(\omega-1)} (\mathcal{V}(-\omega t) - 1)\mathcal{K}\phi\mathcal{K}\psi\right\| = \left\|\left(e^{-\frac{i\xi^2}{2\omega t}} - 1\right)\mathcal{F}(\mathcal{K}\phi\mathcal{K}\psi)\right\|$$

$$\le C t^{\eta - \frac{1}{2}} \|\mathcal{F}(\mathcal{K}\phi\mathcal{K}\psi)\|_{0,1-2\eta} = C t^{\eta - \frac{1}{2}} \|\mathcal{K}\phi\mathcal{K}\psi\|_{1-2\eta,0} \le C t^{\eta - \frac{1}{2}} \|\mathcal{K}\phi\|_{1,0} \|\mathcal{K}\psi\|_{1,0}$$

and

$$\left\|\mathcal{D}_\omega E^{\omega(\omega-1)} \mathcal{K}\psi (\mathcal{V}(-\omega t) - 1)\mathcal{K}\phi\right\| \le C \|\mathcal{K}\psi\|_{2+\frac{2}{\eta}} \left\|\left(e^{-\frac{i\xi^2}{2\omega t}} - 1\right)\mathcal{F}\mathcal{K}\phi\right\|_{2+2\eta}$$

$$\le C t^{\eta - \frac{1}{2}} \|\mathcal{K}\psi\|_{1,0} \|\mathcal{F}\mathcal{K}\phi\|_{0,1} = C t^{\eta - \frac{1}{2}} \|\mathcal{K}\psi\|_{1,0} \|\mathcal{K}\phi\|_{1,0},$$

whence the first estimate of the lemma follows. To prove the second estimate we apply the Hölder inequality and the Sobolev embedding theorem:

$$\left\|\mathcal{D}_\omega E^{\omega(\omega-1)}\left(\mathcal{V}\left(-\omega t\right)-1\right)\mathcal{K}\phi\mathcal{K}\psi\right\|_\infty \le C\left\|\mathcal{F}\left(\mathcal{K}\phi\mathcal{K}\psi\right)\right\|_1$$
$$\le C\left\|\mathcal{K}\phi\mathcal{K}\psi\right\|_{2,0}^{\eta}\left\|\mathcal{K}\phi\mathcal{K}\psi\right\|_{1,0}^{1-\eta} \le C\left\|\mathcal{K}\phi\right\|_{2,0}^{\eta}\left\|\mathcal{K}\psi\right\|_{2,0}^{\eta}\left\|\mathcal{K}\phi\right\|_{1,0}^{1-\eta}\left\|\mathcal{K}\psi\right\|_{1,0}^{1-\eta}.$$

In the same manner we have

$$\left\|\mathcal{D}_\omega E^{\omega(\omega-1)}\mathcal{K}\psi\left(\mathcal{V}\left(-\omega t\right)-1\right)\mathcal{K}\phi\right\|_\infty \le C\left\|\mathcal{K}\phi\right\|_{2,0}^{\eta}\left\|\mathcal{K}\psi\right\|_{2,0}^{\eta}\left\|\mathcal{K}\phi\right\|_{1,0}^{1-\eta}\left\|\mathcal{K}\psi\right\|_{1,0}^{1-\eta}.$$

Thus we have the second estimate of the lemma. Lemma 2.6 is proved. □

3. Proof of Theorem 1.1. We prove Theorem 1.1 via the contraction mapping principle in the functional space $\mathbf{X}$. For any $v \in \mathbf{X}$ with sufficiently small norm $\|v\|_{\mathbf{X}}$, we consider the linearized version of (1.1),

$$\begin{cases} i\partial_t u + \frac{1}{2}\Delta u = \mathcal{N}(v), & (t,x) \in \mathbf{R}\times\mathbf{R}^2, \\ u\left(0,x\right) = u_0(x), & x \in \mathbf{R}^2, \end{cases} \tag{3.1}$$

where $\mathcal{N}(v) = \sum_{j,k=1}^{2}\left(\lambda_{jk}\left(\partial_{x_j}v\right)\left(\partial_{x_k}v\right) + \mu_{jk}\left(\partial_{x_j}\bar{v}\right)\left(\partial_{x_k}\bar{v}\right)\right)$, $\lambda_{jk},\mu_{jk} \in \mathbf{C}$. The linear Cauchy problem (3.1) defines a mapping $u = \mathcal{B}v$. We prove that $\mathcal{B}$ is a contraction mapping from a closed ball in $\mathbf{X}$ with a center at the origin and a radius $C\varepsilon$ into itself. Multiplying both sides of (3.1) by $\mathcal{K}\Gamma^{\alpha+\delta}$, where $\mathcal{K} = \mathcal{F}\mathcal{M}\mathcal{U}\left(-t\right)$, we get

$$\mathcal{L}\mathcal{K}\Gamma^{\alpha+\delta}u = \mathcal{K}\widetilde{\Gamma}^{\alpha+\delta}\mathcal{N} = \frac{1}{it}\sum_{\beta\le\alpha}\sum_{\gamma\le\delta}C_\alpha^\beta C_\delta^\gamma\sum_{j,k=1}^{2}\left(\lambda_{jk}E\left(\mathcal{K}\widetilde{\Gamma}^{\delta-\gamma}f\right)\mathcal{K}\Gamma^\gamma g\right.$$
$$\left.+\mu_{jk}\overline{E}^3\left(\overline{\mathcal{K}\widetilde{\Gamma}^{\delta-\gamma}f}\right)\overline{\mathcal{K}\Gamma^\gamma g}\right), \tag{3.2}$$

where $\mathcal{L} = i\partial_t + \frac{1}{2t^2}\Delta_\xi$, $f = \widetilde{\Gamma}^{\alpha-\beta}\partial_{x_j}v$, $g = \Gamma^\beta\partial_{x_k}v$, $C_\alpha^\beta = \frac{\alpha!}{(\alpha-\beta)!\beta!}$, $|\delta| \le 3$. Applying operator $\mathcal{V}\left(t\right)\varphi = \mathcal{F}^{-1}\mathcal{M}\left(t\right)\mathcal{F}\varphi$ to both sides of (3.2), we obtain by virtue of identity (2.5)

$$i\partial_t\mathcal{V}\left(-t\right)\mathcal{K}\Gamma^{\alpha+\delta}u\left(t\right)$$
$$= \frac{1}{it}\sum_{\gamma\le\delta}\sum_{\beta\le\alpha}C_\delta^\gamma C_\alpha^\beta\sum_{j,k=1}^{2}\mathcal{V}\left(-t\right)\left(\lambda_{jk}E\left(\mathcal{K}\widetilde{\Gamma}^{\delta-\gamma}f\right)\mathcal{K}\Gamma^\gamma g\right.$$
$$\left.+\mu_{jk}\overline{E}^3\left(\overline{\mathcal{K}\widetilde{\Gamma}^{\delta-\gamma}f}\right)\overline{\mathcal{K}\Gamma^\gamma g}\right)$$
$$= \frac{1}{it}\sum_{\gamma\le\delta}\sum_{\beta\le\alpha}C_\delta^\gamma C_\alpha^\beta\sum_{j,k=1}^{2}\left(\lambda_{jk}\mathcal{D}_2E^2\mathcal{V}\left(-2t\right)\left(\mathcal{K}\widetilde{\Gamma}^{\delta-\gamma}f\right)\mathcal{K}\Gamma^\gamma g\right.$$
$$\left.+\mu_{jk}\mathcal{D}_{-2}E^6\mathcal{V}\left(2t\right)\left(\overline{\mathcal{K}\widetilde{\Gamma}^{\delta-\gamma}f}\right)\overline{\mathcal{K}\Gamma^\gamma g}\right). \tag{3.3}$$

Then we write the identity

$$\mathcal{V}\left(-2t\right)\left(\mathcal{K}\widetilde{\Gamma}^{\delta-\gamma}f\right)\mathcal{K}\Gamma^\gamma g$$
$$= \left(\mathcal{V}\left(-2t\right)\mathcal{K}\widetilde{\Gamma}^{\delta-\gamma}f\right)\mathcal{V}\left(-2t\right)\mathcal{K}\Gamma^\gamma g - \left(\mathcal{K}\Gamma^\gamma g\right)\left(\mathcal{V}\left(-2t\right)-1\right)\left(\mathcal{K}\widetilde{\Gamma}^{\delta-\gamma}f\right)$$
$$-\left(\mathcal{V}\left(-2t\right)\mathcal{K}\widetilde{\Gamma}^{\delta-\gamma}f\right)\left(\mathcal{V}\left(-2t\right)-1\right)\mathcal{K}\Gamma^\gamma g + \left(\mathcal{V}\left(-2t\right)-1\right)\left(\mathcal{K}\widetilde{\Gamma}^{\delta-\gamma}f\right)\mathcal{K}\Gamma^\gamma g.$$

By Lemma 2.6 we have

$$\begin{aligned}&\left\|\mathcal{D}_2E^2\left(\mathcal{V}\left(-2t\right)-1\right)\left(\mathcal{K}\widetilde{\Gamma}^{\delta-\gamma}f\right)\mathcal{K}\Gamma^{\gamma}g\right\|\\&+\left\|\mathcal{D}_2E^2\left(\mathcal{K}\Gamma^{\gamma}g\right)\left(\mathcal{V}\left(-2t\right)-1\right)\mathcal{K}\widetilde{\Gamma}^{\delta-\gamma}f\right\|\\&\le C\left|t\right|^{\eta-\frac{1}{2}}\left(\sum_{|\sigma|\le1}\left\|\mathcal{J}^{\sigma}\widetilde{\Gamma}^{\delta-\gamma}f\right\|\right)\left(\sum_{|\sigma|\le1}\left\|\mathcal{J}^{\sigma}\Gamma^{\gamma}g\right\|\right).\end{aligned}\tag{3.4}$$

Thus we may rewrite (3.3) in the form

$$\begin{aligned}&i\partial_t\mathcal{V}\left(-t\right)\mathcal{K}\Gamma^{\alpha+\delta}u\left(t\right)\\&=\frac{1}{it}\sum_{\gamma\le\delta}\sum_{\beta\le\alpha}C_{\delta}^{\gamma}C_{\alpha}^{\beta}\sum_{j,k=1}^{2}\left(\lambda_{jk}\mathcal{D}_2E^2\left(\mathcal{V}\left(-t\right)\mathcal{K}\widetilde{\Gamma}^{\delta-\gamma}f\right)\mathcal{V}\left(-t\right)\mathcal{K}\Gamma^{\gamma}g\right.\\&\left.+\mu_{jk}\mathcal{D}_{-2}E^6\left(\overline{\mathcal{V}\left(-t\right)\mathcal{K}\widetilde{\Gamma}^{\delta-\gamma}f}\right)\overline{\mathcal{V}\left(-t\right)\mathcal{K}\Gamma^{\gamma}g}\right)+R_1\left(t\right),\end{aligned}\tag{3.5}$$

where the remainder term $R_1\left(t\right)$ can be estimated by virtue of (3.4) and Lemmas 2.1–2.4 as follows:

$$\begin{aligned}&\sum_{\alpha}\frac{b^{|\alpha|}}{\alpha!}\left\|R_1\left(t\right)\right\|\\&\le C\left|t\right|^{\eta-\frac{3}{2}}\sum_{\gamma\le\delta}\sum_{\alpha}\sum_{\beta\le\alpha}C_{\delta}^{\gamma}C_{\alpha}^{\beta}\frac{b^{\alpha}}{\alpha!}\left(\sum_{|\sigma|\le1}\left\|\mathcal{J}^{\sigma}\widetilde{\Gamma}^{\delta-\gamma}f\right\|\right)\sum_{|\sigma|\le1}\left\|\mathcal{J}^{\sigma}\Gamma^{\gamma}g\right\|\\&\le C\varepsilon\left|t\right|^{\eta-\frac{3}{2}}\sum_{|\sigma|\le1}\left\|\Theta^{\sigma}v\left(t\right)\right\|_{\mathbf{A}^3(t)}.\end{aligned}\tag{3.6}$$

Similarly, we have for $|\delta|\le1$

$$\begin{aligned}&\sum_{\alpha}\frac{b^{|\alpha|}}{\alpha!}\left\|R_1\left(t\right)\right\|_{\infty}\\&\le\frac{C}{t}\sum_{\alpha}\sum_{\beta\le\alpha}\sum_{\gamma\le\delta}C_{\delta}^{\gamma}C_{\alpha}^{\beta}\frac{b^{\alpha}}{\alpha!}\left(\sum_{|\sigma|\le2}\left\|\mathcal{J}^{\sigma}\widetilde{\Gamma}^{\delta-\gamma}f\right\|^{\eta}\right)\sum_{|\sigma|\le1}\left\|\mathcal{J}^{\sigma}\widetilde{\Gamma}^{\delta-\gamma}f\right\|^{1-\eta}\\&\times\left(\sum_{|\sigma|\le1}\left\|\mathcal{J}^{\sigma}\Gamma^{\gamma}g\right\|^{1-\eta}\right)\sum_{|\sigma|\le2}\left\|\mathcal{J}^{\sigma}\Gamma^{\gamma}g\right\|^{\eta}\le C\varepsilon^2\left|t\right|^{\eta-1}.\end{aligned}\tag{3.7}$$

Since $\mathcal{V}\left(-t\right)\mathcal{K}=\overline{\mathcal{F}^{-1}\mathcal{M}\mathcal{F}}\mathcal{F}\mathcal{M}\mathcal{U}\left(-t\right)=\mathcal{F}\mathcal{U}\left(-t\right)$, we have $i\xi_j\mathcal{V}\left(-t\right)\mathcal{K}=\mathcal{V}\left(-t\right)\mathcal{K}\partial_{x_j}$. Hence by (3.5) we get

$$\begin{aligned}&i\partial_t\mathcal{V}\left(-t\right)\mathcal{K}\Gamma^{\alpha+\delta}u\left(t\right)\\&=\frac{1}{t}\sum_{\gamma\le\delta}\sum_{\beta\le\alpha}C_{\delta}^{\gamma}C_{\alpha}^{\beta}\sum_{j,k=1}^{2}\left(\lambda_{jk}\mathcal{D}_2E^2\xi_j\left(\mathcal{V}\left(-t\right)\mathcal{K}\partial_{x_j}^{-1}\widetilde{\Gamma}^{\delta-\gamma}f\right)\mathcal{V}\left(-t\right)\mathcal{K}\Gamma^{\gamma}g\right.\\&\left.-\mu_{jk}\mathcal{D}_{-2}E^6\xi_j\left(\overline{\mathcal{V}\left(-t\right)\mathcal{K}\partial_{x_j}^{-1}\widetilde{\Gamma}^{\delta-\gamma}f}\right)\overline{\mathcal{V}\left(-t\right)\mathcal{K}\Gamma^{\gamma}g}\right)+R_1\left(t\right).\end{aligned}\tag{3.8}$$

If $|\delta-\gamma|<|\gamma|$ we exchange f and g in the right-hand side of (3.8). By virtue of the equality $E^\omega=\left(1+\frac{it}{2}\omega\xi^2\right)^{-1}\partial_t\left(tE^\omega\right)$ we obtain the identity

$$\frac{\phi}{t}E^\omega=\partial_t\left(\frac{\phi E^\omega}{1+\frac{it}{2}\omega\xi^2}\right)-\frac{E^\omega\partial_t\phi}{1+\frac{it}{2}\omega\xi^2}+\frac{1+it\omega\xi^2}{t\left(1+\frac{it}{2}\omega\xi^2\right)^2}\phi E^\omega.$$

Therefore, we get from (3.8)

$$i\partial_t\Psi=R_2, \tag{3.9}$$

where

$$\begin{aligned}\Psi&=\mathcal{V}\left(-t\right)\mathcal{K}\Gamma^{\alpha+\delta}u\left(t\right)\\&\quad+\sum_{\gamma\le\delta}\sum_{\beta\le\alpha}C_\delta^\gamma C_\alpha^\beta\sum_{j,k=1}^{2}\left(\lambda_{jk}\mathcal{D}_2\frac{\xi_jE^2}{1+it\xi^2}\left(\mathcal{V}\left(-t\right)\mathcal{K}\partial_{x_j}^{-1}\widetilde{\Gamma}^{\delta-\gamma}f\right)\mathcal{V}\left(-t\right)\mathcal{K}\Gamma^\gamma g\right.\\&\qquad\left.+\mu_{jk}\mathcal{D}_{-2}\frac{\xi_jE^6}{1+3it\xi^2}\left(\overline{\mathcal{V}\left(-t\right)\mathcal{K}\partial_{x_j}^{-1}\widetilde{\Gamma}^{\delta-\gamma}f}\right)\overline{\mathcal{V}\left(-t\right)\mathcal{K}\Gamma^\gamma g}\right),\end{aligned}$$

$R_2=R_1+\sum_{j=1}^3 I_j$, and

$$\begin{aligned}I_1&=\frac{1}{it}\sum_{\gamma\le\delta}\sum_{\beta\le\alpha}C_\delta^\gamma C_\alpha^\beta\sum_{j,k=1}^{2}\left(\lambda_{jk}\mathcal{D}_2\frac{\left(1+2it\xi^2\right)E^2}{\left(1+it\xi^2\right)^2}\xi_j\left(\mathcal{V}\left(-t\right)\mathcal{K}\partial_{x_j}^{-1}\widetilde{\Gamma}^{\delta-\gamma}f\right)\right.\\&\quad\left.\times\mathcal{V}\left(-t\right)\mathcal{K}\Gamma^\gamma g-\mu_{jk}\mathcal{D}_{-2}\frac{\left(1+6it\xi^2\right)E^6}{\left(1+3it\xi^2\right)^2}\xi_j\left(\overline{\mathcal{V}\left(-t\right)\mathcal{K}\partial_{x_j}^{-1}\widetilde{\Gamma}^{\delta-\gamma}f}\right)\overline{\mathcal{V}\left(-t\right)\mathcal{K}\Gamma^\gamma g}\right),\end{aligned}$$

$$\begin{aligned}I_2&=-\frac{1}{it}\sum_{\gamma\le\delta}\sum_{\beta\le\alpha}C_\delta^\gamma C_\alpha^\beta\sum_{j,k=1}^{2}\lambda_{jk}\mathcal{D}_2\frac{\xi_jE^2}{1+it\xi^2}\left(\left(\partial_t\mathcal{V}\left(-t\right)\mathcal{K}\partial_{x_j}^{-1}\widetilde{\Gamma}^{\delta-\gamma}f\right)\times\mathcal{V}\left(-t\right)\mathcal{K}\Gamma^\gamma g\right.\\&\qquad\left.+\left(\mathcal{V}\left(-t\right)\mathcal{K}\partial_{x_j}^{-1}\widetilde{\Gamma}^{\delta-\gamma}f\right)\partial_t\mathcal{V}\left(-t\right)\mathcal{K}\Gamma^\gamma g\right),\end{aligned}$$

and

$$\begin{aligned}I_3&=-\frac{1}{it}\sum_{\gamma\le\delta}\sum_{\beta\le\alpha}C_\delta^\gamma C_\alpha^\beta\sum_{j,k=1}^{2}\mu_{jk}\mathcal{D}_{-2}\frac{\xi_jE^6}{1+3it\xi^2}\left(\left(\overline{\partial_t\mathcal{V}\left(-t\right)\mathcal{K}\partial_{x_j}^{-1}\widetilde{\Gamma}^{\delta-\gamma}f}\right)\right.\\&\quad\left.\times\overline{\mathcal{V}\left(-t\right)\mathcal{K}\Gamma^\gamma g}+\left(\overline{\mathcal{V}\left(-t\right)\mathcal{K}\partial_{x_j}^{-1}\widetilde{\Gamma}^{\delta-\gamma}f}\right)\overline{\partial_t\mathcal{V}\left(-t\right)\mathcal{K}\Gamma^\gamma g}\right).\end{aligned}$$

By the Hölder's inequality, the identities $\mathcal{V}\left(-t\right)\mathcal{K}=\mathcal{F}\mathcal{U}\left(-t\right)$, $\mathcal{J}=\mathcal{U}\left(t\right)x\mathcal{U}\left(-t\right)$, and Lemmas 2.1–2.4 we get the estimates

$$\begin{aligned}&\sum_\alpha\frac{b^{|\alpha|}}{\alpha!}\left\|I_1\left(t\right)\right\|\\&\le C\left|t\right|^{-\frac{3}{2}}\sum_\alpha\sum_{\beta\le\alpha}C_\alpha^\beta\frac{b^{|\alpha|}}{\alpha!}\sum_{j,k=1}^{2}\left(\sum_{|\delta|\le3}\left\|\partial_{x_j}^{-1}\widetilde{\Gamma}^\delta f\right\|\right)\sum_{|\gamma|\le1}\left\|\mathcal{V}\left(-t\right)\mathcal{K}\Gamma^\gamma g\right\|_\infty\end{aligned}$$

$$
\begin{aligned}
&\le C\left|t\right|^{-\frac{3}{2}}\sum_{\alpha}\sum_{\beta\le\alpha}C_{\alpha}^{\beta}\frac{b^{|\alpha|}}{\alpha!}\left(\sum_{|\delta|\le3}\left\|\partial_{x_j}^{-1}\widetilde{\Gamma}^{\delta}f\right\|\right)\\
&\times\sum_{|\gamma|\le1}\left(\sum_{|\sigma|\le1}\left\|\mathcal{J}^{\sigma}\Gamma^{\gamma}g\right\|^{1-\eta}\right)\sum_{|\sigma|\le2}\left\|\mathcal{J}^{\sigma}\Gamma^{\gamma}g\right\|^{\eta} \qquad (3.10)\\
&\le C\varepsilon^{2}\left|t\right|^{2\eta-\frac{3}{2}}
\end{aligned}
$$

for $|\delta|\le3$. In the same way we obtain

$$
\begin{aligned}
&\sum_{\alpha}\frac{b^{|\alpha|}}{\alpha!}\left(\left\|I_2\left(t\right)\right\|+\left\|I_3\left(t\right)\right\|\right)\\
&\le C\left|t\right|^{-\frac{1}{2}}\sum_{\alpha}\sum_{\beta\le\alpha}\sum_{|\gamma|\le1}C_{\delta}^{\gamma}C_{\alpha}^{\beta}\frac{b^{|\alpha|}}{\alpha!}\sum_{j,k=1}^{2}\left(\left\|\partial_t\mathcal{V}\left(-t\right)\mathcal{K}\partial_{x_j}^{-1}\widetilde{\Gamma}^{\delta-\gamma}f\right\|\times\left\|\mathcal{V}\left(-t\right)\mathcal{K}\Gamma^{\gamma}g\right\|_{\infty}\right.\\
&\qquad\left.+\left\|\partial_t\mathcal{V}\left(-t\right)\mathcal{K}\Gamma^{\gamma}g\right\|_{\infty}\left\|\mathcal{V}\left(-t\right)\mathcal{K}\partial_{x_j}^{-1}\widetilde{\Gamma}^{\delta-\gamma}f\right\|\right)\\
&\le C\varepsilon\left|t\right|^{2\eta-\frac{1}{2}}\sum_{\alpha}\frac{b^{|\alpha|}}{\alpha!}\left(\sum_{|\delta|\le3}\left\|\partial_t\mathcal{V}\left(-t\right)\mathcal{K}\widetilde{\Gamma}^{\alpha+\delta}v\right\|\right.\\
&\qquad\left.+\sum_{|\gamma|\le1}\left\|\partial_t\mathcal{V}\left(-t\right)\mathcal{K}\Gamma^{\alpha+\gamma}\nabla v\right\|_{\infty}\right). \qquad (3.11)
\end{aligned}
$$

In view of (3.6), (3.10), and (3.11) we have the estimate

$$
\begin{aligned}
&\sum_{\alpha}\frac{b^{|\alpha|}}{\alpha!}\left\|R_2\left(t\right)\right\|\\
&\le C\varepsilon\left|t\right|^{\eta-\frac{3}{2}}\left\|v\right\|_{\mathbf{A}^4(t)}+C\varepsilon^{2}\left|t\right|^{2\eta-\frac{3}{2}}\\
&\quad+C\varepsilon\left|t\right|^{2\eta-\frac{1}{2}}\sum_{|\delta|\le3}\sum_{\alpha}\frac{b^{|\alpha|}}{\alpha!}\left\|\partial_t\mathcal{V}\left(-t\right)\mathcal{K}\Gamma^{\alpha+\delta}v\right\|\\
&\quad+C\varepsilon\left|t\right|^{2\eta-\frac{1}{2}}\sum_{|\gamma|\le1}\sum_{\alpha}\frac{b^{|\alpha|}}{\alpha!}\left\|\partial_t\mathcal{V}\left(-t\right)\mathcal{K}\Gamma^{\alpha+\gamma}\nabla v\right\|_{\infty}. \qquad (3.12)
\end{aligned}
$$

Multiplying both sides of (3.9) by $\overline{\Psi\left(t\right)}$, integrating with respect to the space variables and taking the imaginary part of the result, we obtain the inequality $\frac{d}{dt}\left\|\Psi\left(t\right)\right\|\le\left\|R_2\left(t\right)\right\|$, whence

$$
\frac{d}{dt}\sum_{\alpha}\frac{b^{|\alpha|}}{\alpha!}\left\|\Psi\left(t\right)\right\|-b'\sum_{\alpha}\sum_{|\delta|=1}\frac{b^{|\alpha-\delta|}}{(\alpha-\delta)!}\left\|\Psi\left(t\right)\right\|\le\sum_{\alpha}\frac{b^{|\alpha|}}{\alpha!}\left\|R_2\left(t\right)\right\|.
$$

Then integration with respect to t in view of (3.12) yields

$$\begin{aligned}
&\|u(t)\|_{\mathbf{A}^3(t)} + C\sum_{|\sigma|=1}\int_1^t \|\Gamma^\sigma u(\tau)\|_{\mathbf{A}^3(\tau)}\,|b'|\,d\tau \\
&\le C\varepsilon + C\varepsilon^2 + \frac{C}{\sqrt{t}}\,\|v(t)\|^{2-\eta}_{\mathbf{A}^3(t)} \sum_{|\gamma|\le 1}\|\mathcal{J}^\gamma v\|^\eta_{\mathbf{A}^2(t)} \\
&\quad + C\int_1^t \|v(\tau)\|^{1-\eta}_{\mathbf{A}^3(\tau)}\,\|v(\tau)\|^{1-\eta}_{\mathbf{A}^4(\tau)} \left(\|v(\tau)\|^\eta_{\mathbf{A}^3(\tau)} \sum_{|\gamma|\le 1}\|\mathcal{J}^\gamma v\|^\eta_{\mathbf{A}^2(\tau)} \right. \\
&\qquad\qquad \left. + \|v(\tau)\|^\eta_{\mathbf{A}^4(\tau)} \sum_{|\gamma|\le 1}\|\mathcal{J}^\gamma v\|^\eta_{\mathbf{A}^3(\tau)} \right) \frac{|b'|\,d\tau}{\sqrt{\tau}} \\
&\le C\varepsilon + C\varepsilon^2, \qquad (3.13)
\end{aligned}$$

where we used the estimate

$$\|u(t)\|_{\mathbf{A}^3(t)} + C\sum_{|\alpha|=1}\int_0^t \|\Gamma^\sigma u(t)\|_{\mathbf{A}^3(t)}\,|b'|\,dt \le C\varepsilon + C\varepsilon^2$$

for $t \le 1$, which follows from the classical energy method. Applying the operator $\mathcal{J}_{x_l}$ to both sides of (3.1), we get

$$\begin{aligned}
&\left(i\partial_t + \frac{1}{2}\Delta\right)\mathcal{J}_{x_l}\Gamma^{\alpha+\delta}u \\
&= \sum_{\gamma\le\delta}\sum_{\beta\le\alpha} C^\gamma_\delta C^\beta_\alpha \sum_{j,k=1}^2 \left(\lambda_{jk}\left(\widetilde{\Gamma}^{\delta-\gamma}f\right)\mathcal{J}_{x_l}\Gamma^\gamma g + it\lambda_{jk}\left(\partial_{x_l}\widetilde{\Gamma}^{\delta-\gamma}f\right)\Gamma^\gamma g \right. \\
&\qquad \left. + \mu_{jk}\overline{\left(\widetilde{\Gamma}^{\delta-\gamma}f\right)\mathcal{J}_{x_l}\Gamma^\gamma g} - it\mu_{jk}\overline{\left(\partial_{x_l}\widetilde{\Gamma}^{\delta-\gamma}f\right)\Gamma^\gamma g}\right);
\end{aligned}$$

hence by the classical energy method and via inequality $\left\|\left(\partial_{x_j}\phi\right)\left(\partial_{x_k}\psi\right)\right\| \le \frac{C}{t}\,\|\Theta\phi\|\,\|\Theta\psi\|$, we obtain

$$\begin{aligned}
&\sum_\alpha \frac{b^{|\alpha|}}{\alpha!}\frac{d}{dt}\left\|\mathcal{J}_{x_l}\Gamma^{\alpha+\delta}u(t)\right\| \\
&\le C\left(\|v\|_{\mathbf{A}^3(t)} + \sum_{|\sigma|\le 1}\|\Theta^\sigma v\|_{\mathbf{A}^2(t)}\right)\sum_{|\gamma|\le 1}\|\Theta^\gamma v\|_{\mathbf{A}^2(t)} \le C\varepsilon^2
\end{aligned}$$

for $|\delta| \le 2$. Multiplying both sides of the last inequality by $t^{-1-\eta}$ and integrating with respect to t, we have

$$\begin{aligned}
&t^{-1-\eta}\sum_{|\gamma|=1}\|\mathcal{J}^\gamma u(t)\|_{\mathbf{A}^2(t)} + C\sum_{|\gamma|=1,|\sigma|=1}\int_1^t \|\Gamma^\sigma\mathcal{J}^\gamma u(t)\|_{\mathbf{A}^2(t)}\,\frac{|b'|\,dt}{t^{1+\eta}} \\
&\qquad\qquad \le C\varepsilon + C\varepsilon^2, \qquad (3.14)
\end{aligned}$$

where we used the estimate

$$\sum_{|\gamma|=1}\|\mathcal{J}^\gamma u(t)\|_{\mathbf{A}^2(t)} + C\sum_{|\gamma|=|\delta|=1}\int_0^t \left\|\Gamma^\delta\mathcal{J}^\gamma u(t)\right\|_{\mathbf{A}^2(t)}\,|b'|\,dt \le C\varepsilon + C\varepsilon^2$$

for $t \le 1$. By the identity $\mathcal{P}u = \mathcal{Q}u + 2it\mathcal{N}(v)$ we see that

$$\|\mathcal{Q}u\|_{\mathbf{A}^m(t)} \le \|\mathcal{P}u\|_{\mathbf{A}^m(t)} + C\|\mathcal{Q}v\|_{\mathbf{A}^m(t)}\|\mathcal{Q}v\|_{\mathbf{A}^m(t)} \le \|\mathcal{P}u\|_{\mathbf{A}^m(t)} + C\varepsilon^2,$$

where $m \ge 2$. Therefore, by virtue of (3.13) and (3.14) we get

$$\begin{aligned}
&\|u(t)\|_{\mathbf{A}^3(t)} + t^{-1-\eta}\sum_{|\gamma|=1}\|\mathcal{J}^\gamma u(t)\|_{\mathbf{A}^2(t)} \\
&+C\sum_{|\sigma|=1}\int_1^t \|\Theta^\sigma u(t)\|_{\mathbf{A}^3(t)}|b'|\,dt \\
&+C\sum_{|\sigma|=1,|\gamma|\le 1}\int_1^t \|\Theta^\sigma \mathcal{J}^\gamma u(t)\|_{\mathbf{A}^2(t)}\frac{|b'|\,dt}{t^{1+\eta}} \\
&\le C\varepsilon + C\varepsilon^2. \qquad (3.15)
\end{aligned}$$

By (3.3) via Lemma 2.6 we have

$$\sum_{|\sigma|\le 1}\sum_\alpha \frac{b^{|\alpha|}}{\alpha!}\left\|\partial_t\mathcal{V}(-t)\mathcal{K}\Gamma^{\alpha+\sigma}\nabla u(t)\right\|_\infty \le C\varepsilon^2 t^{2\eta-1}$$

and

$$\sum_{|\delta|\le 3}\sum_\alpha \int_1^\infty \frac{b^{|\alpha|}}{\alpha!}\left\|\partial_t\mathcal{V}(-t)\mathcal{K}\Gamma^{\alpha+\delta}u(t)\right\|\frac{dt}{t^{\frac{1}{2}-2\eta}} \le C\varepsilon\int_1^\infty t^{3\eta-\frac{3}{2}}\|v(t)\|_{\mathbf{A}^3(t)}\,dt \le C\varepsilon^2.$$

We also have by a classical energy method

$$\sum_{|\sigma|\le 1}\sum_\alpha \frac{b^{|\alpha|}}{\alpha!}\left\|\partial_t\mathcal{V}(-t)\mathcal{K}\Gamma^{\alpha+\sigma}\nabla u(t)\right\|_\infty \le C\varepsilon^2$$

and

$$\sum_{|\delta|\le 3}\sum_\alpha \int_0^t \frac{b^{|\alpha|}}{\alpha!}\left\|\partial_t\mathcal{V}(-t)\mathcal{K}\Gamma^{\alpha+\delta}u(t)\right\|dt \le C\varepsilon^2$$

for $t \le 1$. In the same way as in the proof of (3.15) we find that $\|\mathcal{B}v_1 - \mathcal{B}v_2\|_{\mathbf{X}} \le \frac{1}{2}\|v_1 - v_2\|_{\mathbf{X}}$. Therefore, (3.15) implies that $\mathcal{B}$ is a contraction mapping from $\mathbf{X}$ into itself. Hence there exists a unique solution u of the Cauchy problem (1.1) such that $u \in \mathbf{X}$. In particular we have $u \in \mathbf{A}^3(t)$. Next we show the existence of the usual scattering states. From (3.9) with $\alpha = 0$, $\delta = 0$ we have

$$\begin{aligned}
i\partial_t\mathcal{V}(-t)\mathcal{K}u(t) - \partial_t\sum_{\beta\le\alpha}C_\alpha^\beta\sum_{j,k=1}^2\Bigg(&\lambda_{jk}\mathcal{D}_2\frac{\xi_j E^2}{1+it\xi^2}\left(\mathcal{V}(-t)\mathcal{K}u\right)\mathcal{V}(-t)\mathcal{K}\partial_{x_k}u \\
&+\mu_{jk}\mathcal{D}_{-2}\frac{\xi_j E^6}{1+3it\xi^2}\left(\overline{\mathcal{V}(-t)\mathcal{K}u}\right)\overline{\mathcal{V}(-t)\mathcal{K}\partial_{x_k}u}\Bigg) \\
&= R_3(t),
\end{aligned}$$

where $\|R_3(t)\| \le C\varepsilon^2|t|^{4\eta-\frac{3}{2}}$. Integrating this equation with respect to t, we obtain

$$\|\mathcal{V}(-t)\mathcal{K}u(t) - \mathcal{V}(-s)\mathcal{K}u(s)\| \le C\varepsilon^2|s|^{4\eta-\frac{1}{2}} \qquad (3.16)$$

for $t > s > 0$. Here we used the estimate

$$\begin{aligned}&\left\|\mathcal{D}_2 \frac{\xi_j E^2}{1+it\xi^2}\left(\mathcal{V}(-t)\mathcal{K}u\right)\mathcal{V}(-t)\mathcal{K}\partial_{x_k}u\right\| \\ &+\left\|\mathcal{D}_{-2}\frac{\xi_j E^6}{1+3it\xi^2}\left(\overline{\mathcal{V}(-t)\mathcal{K}u}\right)\overline{\mathcal{V}(-t)\mathcal{K}\partial_{x_k}u}\right\| \\ &\le C|t|^{-\frac{1}{2}}\|u\|\sum_{|\sigma|\le 1}\|\Theta^\sigma u\|^{1-\eta}\sum_{|\sigma|\le 1,|\gamma|\le 1}\|\mathcal{J}^\gamma\Theta^\sigma u\|^\eta \le C\varepsilon^2|s|^{4\eta-\frac{1}{2}}.\end{aligned}$$

Since $\mathcal{V}(-t)\mathcal{K} = \mathcal{F}\mathcal{U}(-t)$, from the inequality (3.16) we find that there exists a unique final state $u_+ \in \mathbf{L}^2$ such that $\|u(t) - \mathcal{U}(t)u_+\| \le C\varepsilon^2|t|^{\eta-\frac{1}{2}}$. This completes the proof of Theorem 1.1.

Acknowledgment. The authors would like to thank the referee for useful suggestions.

REFERENCES

[1] H. Aikawa, *Infinite order Sobolev spaces, analytic continuation and polynomial expansions*, Complex Variables Theory Appl., 18 (1992), pp. 253–266.
[2] H. Aikawa, N. Hayashi, and S. Saitoh, *Isometrical identities for the Bergman and the Szegö spaces on a sector*, J. Math. Soc. Japan, 43 (1991), pp. 195–201.
[3] S. Cohn, *Resonance and long time existence for the quadratic semilinear Schrödinger equation*, Comm. Pure Appl. Math., 45 (1992), pp. 973–1001.
[4] S. Cohn, *Global existence for the nonresonant Schrödinger equation in two space dimensions*, Canad. Appl. Math. Quart., 2 (1994), pp. 247–282.
[5] J. Ginibre and N. Hayashi, *Almost global existence of small solutions to quadratic nonlinear Schrödinger equations in three space dimensions*, Math. Z., 219 (1995), pp. 119–140.
[6] N. Hayashi, *Global existence of small solutions to quadratic nonlinear Schrödinger equations*, Comm. Partial Differential Equations, 18 (1993), pp. 1109–1124.
[7] N. Hayashi, *Global and almost global solutions to quadratic nonlinear Schrödinger equations with small initial data*, Dynam. Contin. Discrete Impuls. Systems, 2 (1996), pp. 109–129.
[8] N. Hayashi and H. Hirata, *Global existence of small solutions to nonlinear Schrödinger equations*, Nonlinear Anal., 31 (1998), pp. 671–685.
[9] N. Hayashi and K. Kato, *Global existence of small analytic solutions to Schrödinger equations with quadratic nonlinearity*, Comm. Partial Differential Equations, 22 (1997), pp. 773–798.
[10] N. Hayashi and P. I. Naumkin, *On the quadratic nonlinear Schrödinger equation in three space dimensions*, Internat. Math. Res. Notices, 3 (2000), pp. 115–132.
[11] N. Hayashi, C. Miao, and P. I. Naumkin, *Global existence of small solutions to the generalized derivative nonlinear Schrödinger equation*, Asymptot. Anal., 21 (1999), pp. 133–147.
[12] S. Klainerman, *Long-time behavior of solutions to nonlinear evolution equations*, Arch. Rational Mech. Anal., 78 (1982), pp. 73–98.
[13] S. Klainerman and G. Ponce, *Global, small amplitude solutions to nonlinear evolution equations*, Comm. Pure Appl. Math., 36 (1983), pp. 133–141.
[14] T. Ozawa, *Remarks on quadratic nonlinear Schrödinger equations*, Funkcial. Ekvac., 38 (1995), pp. 217–232.
[15] J. Shatah, *Global existence of small solutions to nonlinear evolution equations*, J. Differential Equations, 46 (1982), pp. 409–425.
[16] J. Shatah, *Normal forms and quadratic nonlinear Klein-Gordon equations*, Comm. Pure Appl. Math., 38 (1985), pp. 685–696.
[17] W. A. Strauss, *Nonlinear scattering theory at low energy*, J. Funct. Anal., 41 (1981), pp. 110–133.